Disentangling Head NeRF for 3D Facial Animation

Mingyuan Shen[1], Qun-Ce Xu[1], Tai-Jiang Mu[1]

[1]BNRist, Tsinghua University

Abstract In recent years, significant advancements have been made in audio-driven 3D facial animation, unlocking its vast potential in applications spanning gaming, digital humans, and virtual reality. Nevertheless, existing 3D methods for facial animation have fallen short in delivering truly natural and realistic results, often prioritizing lip movements while neglecting facial expressions across other regions. To address this, we propose a disentangling head NeRF for 3D facial animation. Specifically, we train a head NeRF that completely disentangle semantic latent codes of the face including the expression and the intrinsic properties (Identity) and use an well-designed network architecture and loss items to learn generating realistic facial expressions while keeping lip synchronization via a lip sync module. The experiments show that our work outperform the state-of-the-art methods on objective and subjective metrics.

Keywords Facial Animation, NeRF, Talking Head Generation

1 Introduction

The advancement of 3D digital human technology has ignited a surge of interest in the realm of facial animation for talking heads. This burgeoning field has quickly become a focal point of research within the domains of computer graphics and computer vision, offering immense potential for applications across diverse industries, including film production, gaming, mixed reality, and beyond. The primary objective of facial animation is to produce a realistic and expressive video of a person speaking, synchronized with a specific audio clip. This is achieved by harnessing a series of source videos that capture the person's facial expressions and movements, creating a seamless blend between the audio and visual elements.

Presently, there are notable constraints associated with the existing techniques in this domain. The majority of current approaches, as demonstrated in recent research such as works [1, 2], heavily depend on parametric models like BFM [3], 3DMM [4], and FLAME [5]. These models are employed to offer foundational insights into the shape and texture of facial structures. However, owing to the restricted expressive capabilities of these intermediary models, they encounter challenges in effectively decouple facial expressions from inherent facial attributes. Consequently, this limitation leads to a loss of information and misalignment between the facial expressions and lip movements. Recently, Neural Radiance Field (NeRF)[6] has shown great promise in 3D object synthesis and rendering for its ability to render high-fidelity images with rich details. AD-NeRF[7] directly maps the audio features to neural radiance fields to edit the talking head, DFRF[2] proposed a differentiable face warping module conditioned on audio signals to synthesize talking head videos fast and data efficiently. The intrinsic characteristics and facial expressions on the face are closely intertwined within the latent space of the NeRF. However, the facial animation procedure has the potential to alter the fundamental attributes of the human face, resulting in inaccuracies and unrealism in the generated video.

In this paper, we propose a disentangling Head Neural Radiance Field which completely disentangles one person's identity, expression, albedo and illumination ensuring that the synchronization of lip movements and the conveyance of expressions driven by speech audio are not only accurate but also profoundly realistic. Without the intermediate model such as FLAME[5] or 3DMM[4], our method generates talking head video in

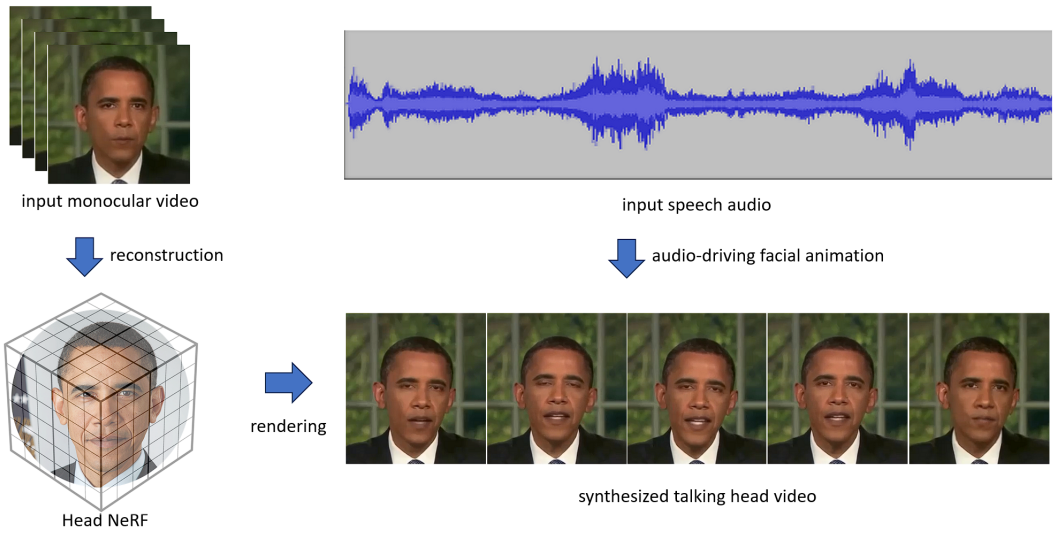

Fig. 1. Illustration of the Facial Animation Task. Our framework takes a short clip of talking head video and a speech audio clip as the input, and trains a dynamic disentangling Head NeRF. A new talking head video are generated from the corresponding input audio.

an end-to-end way. This obviates the potential errors and discrepancies that can arise from these intermediate face representations. Additionally, our method capitalizes on the full spectrum of the video clip, allowing us to capture a personalized facial expression. We enhance this process with a lip synchronization module utilizing contrastive learning which ensures that the lip area is primarily synchronized with the audio clip while the remaining parts of the face are influenced by individualized attributes and talking style.

In summary, our contributions in this work are as follows:

- We propose a comprehensive end-to-end 3D facial animation framework that generates a accurate and highly realistic talking head video clip from an audio clip without any intermediate models.
- We disentangle the attributes including Identity, Expression, Albedo and Illumination of a human face in the latent space and get detailed and accurate expression driven by the speech audio.
- We evaluate the performance of our facial animation model through extensive experimentation on multiple datasets. And the results demonstrate that our method surpasses existing baseline approaches, both in terms of objective metrics and subjective evaluations.

2 Related Work

2.1 Neural Radiance Fields

Neural Radiance Fields (NeRF)[6] technique utilizes a fully-connected neural network to store the geometry and appearance of an object in voxel grids. It allows for implicit modeling of the 3D structure of a specific object without relying on a 3D model. This approach has found success in various aspects of 3D scene.[8, 9]. Initially designed for static object modeling, subsequent works[10] have extended NeRF to dynamic scenes, some of which focused on face representation[11, 12, 13] and editing[8, 14, 15].

Gafni et al.[11] combine NeRF network with a low-dimensional morphable model to provide explicit control over pose and expression, which can be learned from monocular input data only. HeadNeRF[12] pro-

posed a NeRF-based parametric head model that renders high fidelity h ead i mages i n r eal-time, a nd supports directly controlling pose and expressions. Besides, NeRF has also been employed as the fundamental pipeline for talking head synthesis in several works[7, 16, 17, 18], resulting in satisfactory generation results. However, for facial animation task, existing works did not use the prior knowledge of human face and did not seperate the latent space into several semantically meaningful space. Inspired by HeadNeRF[12], our work leverage NeRF for talking video synthesis and use the disentangled latent codes to get an accurate and realistic rendering result.

2.2 2D-based Talking Portrait Synthesis

In the early stage, most works[19, 20, 21, 22, 23]of talking head video generation are based on 2D image techniques such as GAN (Generative adversarial networks) [24] or image-to-image translation[25]. To bridge the gap between audio and face expression, the semantically meaningful information need to be extracted from audio using Automatic Speech Recognition (ASR) models including DeepSpeech[26], Wav2vec [27, 28], and HuBERT[29]. Some works[19, 22] use 2D landmarks as the intermediate face model to encodes the expression, while some works[30, 31, 32] use 3DMM[4].

ATVG[33] devised a cascade GAN approach to generate talking face video that avoids fitting spurious correlations between audiovisual signals that are irrelevant to the speech content. Wav2lip[22] introduced a powerful lip-sync discriminator to morph the lip movements of arbitary identities and proposed a new, rigorous evaluation benchmarks and metrics to measure the accuracy of lip synchronization. LSP[34] leveraged a network that extracts deep audio features and project the features to the target person's speech space. MakeItTalk[19] separated the content and speaker information in the input audio signal and extended it to artistic paintings and cartoon characters. Chen et al.[31] achieves controllable and temporally coherent talking-head videos with natural head movements through modeling the head motion and facial expressions.

As these methods do not acquire the 3D structure of the human face, they inherently lack the capability to facilitate free viewpoint switching and often encounter challenges in maintaining consistency across multiple views. Additionally, the absence of 3D facial information can result in less vivid and realistic expressions, occasionally leading to distortions when compared to models founded on 3D-based representations.

2.3 Audio-driven 3D Talking Head Generation

3D-based audio-driven facial animation methods are capable of generating more realistic and multi-view consistent talking head videos than the 2D-based ones. Early works[35, 36, 37] usually utilized the 3D Morphable Models (3DMM)[4] as an intermediate model to get the prior knowledge of a human head. With the booming of NeRF, several works[7, 16, 17, 18] synthesize the talking head video using NeRF to represent the human face.

AD-NeRF[7] is the first to utilize NeRF for audio-driven 3D facial animation and successfully get the photo-realistic results. RAD-NeRF[16] further improves the representation structure of the face NeRF and achieve real-time facial animation. DFRF.[2] proposes a model that can rapidly generalize to an unseen identity with few training data by conditioning the radiance field on appearance images to learn the face prior. DFA-NeRF[17] takes lip movements features and personalized attributes as two disentangled conditions to learn plausible lip motion, head pose and eye

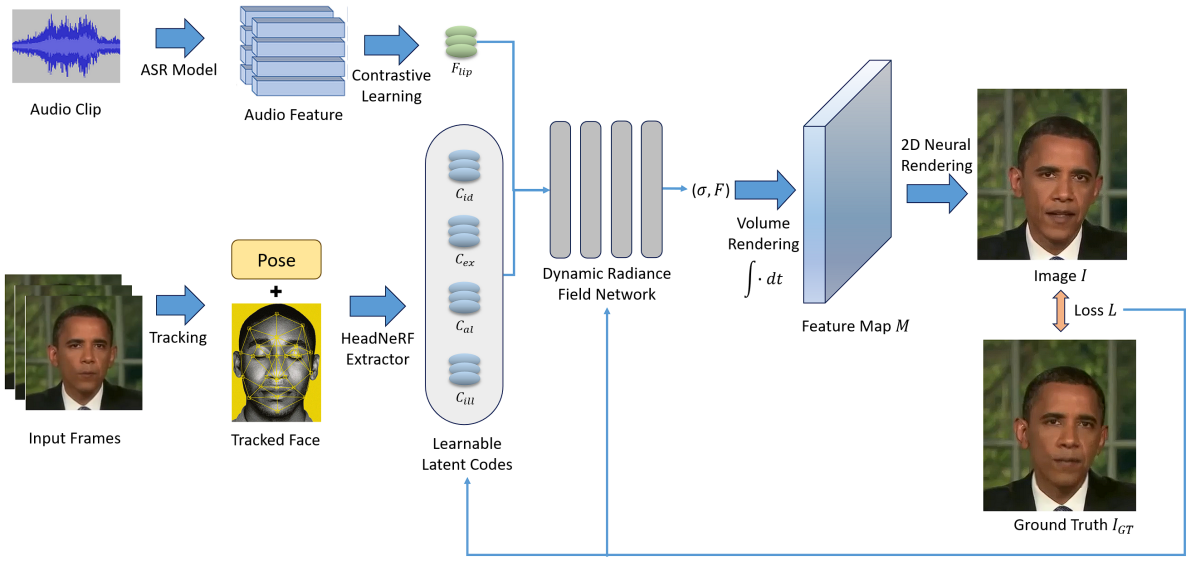

Fig.2. An overview of the architechture of our work. Our method takes several input frames from a monocular talking head video and reconstructs a disentangling head NeRF. The feature input audio are extracted and then convert to lip features with contrastive learning. After volume rendering and 2D nueral rendering, we can calculate the loss items between the result image I and the ground truth I_{GT}.

blink. Geneface[1] enhances the generalizability to out-of-domain audio by learning a motion generator on a large lip-reading corpus, and introduce a domain adaptative post-net to calibrate the result.

Different from these methods, our approach does not rely on a template mesh or an explicit surface representation, such that the error accumulation can be alleviated. Instead, we represent the geometry and appearance implicitly using a NeRF that has disentangling latent space and use volumetric rendering to generate high-fidelity images of the audio-driven talking head.

3 Method

3.1 Overview

In this work, we propose Disentangling NeRF, a novel a disentangling facial Neural Radiance Field model that disentangles one person's Identity, expression, albedo and illumination, so that the movement of the lips and the expressions driven by the speech audio will be accurate and realistic. As shown in Figure 1, the model take a short clip of talking head video and the corresponding speech audio as input, and the final output is the synthesized talking head video rendered by our NeRF. In order to use the audio to drive the expression animation of the face, the audio per-frame feature F_A is extracted by HuBERT[29], the sate-of-the-art speech representation learning model. Meanwhile, the input video frames are processed with Face2Face[38] module to estimate the head pose and the tracked face with landmarks. With these extracted attributes, we can initialize the learnable semantic latent codes C including identity C_{id}, expression C_{ex}, albedo C_{al} and illumination C_{ill} of the image at each time, by which we can render the synthetic talking head video with the proper expression and lip movement using volume rendering. Beside, the loss terms are well designed to ensure the high quality of the reconstruction and the facial animation task(Section 3.3).

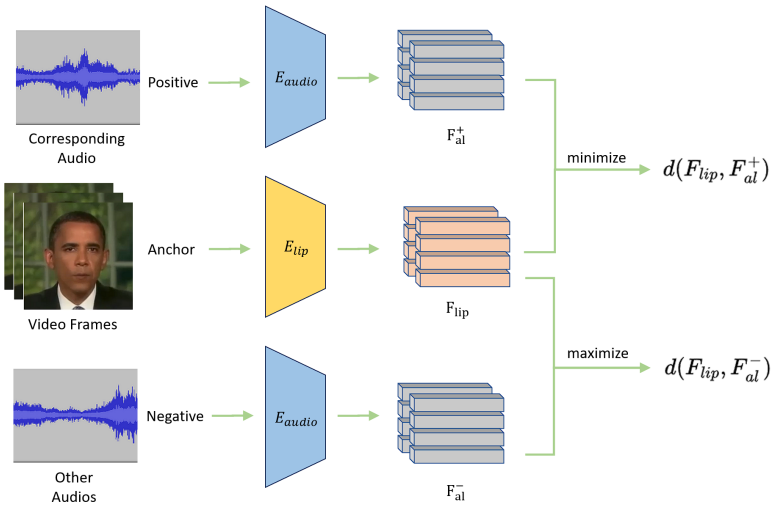

Fig.3. Illustration of the contrastive learning in the Lip Sync module. The features are extracted with the encoders. The positive and negative sample's encoder E_{audio} share the same weights. The goal of the contrastive learning is to maximize the distance between the features of positive sample and the anchor sample, while minimize the distance between the anchor's and the negative's.

3.2 NeRF Model Architecture

We propose a NeRF that can reconstruct a 3D face from several photos from a video clip without camera poses. The model can render a image I_t with several learnable latent codes C, including C_{id}, C_{ex}, C_{al} and C_{ill}, representing identity, expression, albedo and illumination respectively. We first track the motion of head following the face tracking method Face2Face[38] and get the estimated head pose P, and then inversely calculate the camera pose P_{cam} including the rotation matrix R and the translation vector T, assuming that the head is stationary. With the face tracking method, we can get the low dimensional expression parameters of the 3DMM as F_{exp}, which can be used as conditioning for the NeRF model. Besides, with the extracted audio feature F_A, we can learn a lip movement embedding F_{lip} with the Sync Module as the input of volume rendering (See Section 3.4). Inspired by HeadNeRF[12] and some other previous works[39], we predict a high-dimensional intermediate feature map M of the image I using volume rendering, and then use a 2D neural rendering module to get the high-fidelity rendered image.

The radiance field is a function of position coordinate $\mathbf{x}$, viewing direction $\mathbf{v}$ and latent codes C. The volume rendering can be formulated as:

$$D_\theta(\mathbf{x}, \mathbf{v}, P_{cam}, C, F_{lip}) = (\sigma(\mathbf{x}), M(\mathbf{x})) \tag{1}$$

where θ is parameters of the volumetric rendering multi-layer perceptron (MLP), $\sigma(\mathbf{x})$ represents volume density of position $\mathbf{x}$, and $M(\mathbf{x})$ represents the intermediate feature map of the image I which is characterized by 256 dimensions.

After volume rendering, we use a 2D neural rendering module to get the final high-fidelity image I from the intermediate feature map M. Similar to [39] and [12], the neural rendering module consists of several Conv2D and leaky ReLU layer, through which the resolution of the image gradually increases.

3.3 Loss Function

We train our facial animation NeRF on VOCAset[35], a dataset that provides ground truth of 3D facial animation. Our model learn the volumetric rendering network parameters θ, neural rendering network parameters and the latent code C during the training process under the supervision of the following loss terms. We define the total loss as the weighted sum of the loss items below:

$$L_{total} = \lambda_{rec} \cdot L_{rec} + \lambda_{dis} \cdot L_{dis} + \lambda_v \cdot L_v \quad (2)$$

where λ_* is the weights. In our setting, $\lambda_{rec} = 1$, $\lambda_{dis} = 5$, and $\lambda_v = 10$.

Reconstruction Loss The reconstruction loss L_{rec} evaluate the 3D face reconstruction quality of the NeRF model. In this work, it consists of photometric loss and perception loss.

$$L_{rec} = L_{per} + L_{pho} \quad (3)$$

The perception loss[40] L_{per} measures how well the generated image matches the desired image. By comparing the features extracted from the generated and target images at different levels. Perception loss helps guide the optimization process to minimize the discrepancy between the two images, resulting in better image generation, which can be formulated as:

$$L_{per} = \sum_i \|\phi_i(I) - \phi_i(I_{GT})\|^2 \quad (4)$$

where ϕ represents a VGG16 network and i means the i^{th} layer.

The photometric loss L_{pho} measures measures the dissimilarity in the pixel values, color, and texture between the rendered result and the groungtruth image. By utilizing photometric loss, our models can effectively learn to recreate visually accurate and realistic representations of the target human face.

$$L_{pho} = \|m(I) - m(I_{GT})\|^2 \quad (5)$$

where $m(*)$ represents the extraction method[38] of the masked area of human face.

Disentangled Loss In our model, we use contrastive learning method to get the disentangled latent codes of the reconstructed human face. We first pre-train a latent code extractor with the same structure as HeadNeRF[12]. To get a better quality of disentangling of the latent codes, we further use contrastive learning method to train the extractor (the encoders). Specifically, we use the images of the same person with different expressions as negative sample pairs, and the images of different people with the same expression as positive sample pairs so that we can train the expression encoder properly. Similarly, the identity latent code C_{id}'s encoder can be pre-trained with contrastive learning.

With the pre-trained latent code extractor, we can get the referential latent codes $\bar{C_{id}}, \bar{C_{ex}}, \bar{C_{al}}$, and $\bar{C_{ill}}$. In our model, we use the disentangled loss to ensure that the inferred learnable latent codes of the NeRF are near the referential ones. Therefore the disentangled loss can be formulated as:

$$\begin{aligned} L_{dis} = \lambda_{id}\|C_{id} - \bar{C_{id}}\|^2 + \lambda_{ex}\|C_{ex} - \bar{C_{ex}}\|^2 \\ + \lambda_{al}\|C_{al} - \bar{C_{al}}\|^2 + \lambda_{ill}\|C_{ill} - \bar{C_{ill}}\|^2 \end{aligned} \quad (6)$$

where λ_* stands for the weight of each item.

Animation Velocity Loss The facial animation task need to keep time consistency. That is, the extent of change in expression over time should remain roughly constant. In this way, the facial animation result will not suffer from an unnatural sudden change,

thus resulting in a realistic synthesized video. Inspired by VOCA[35], to ensure the temporal consistency, we introduce a velocity loss term L_v as bellow:

$$L_v = \|(I^j - I^{j-1}) - (I^j_{GT} - I^{j-1}_{GT})\|_2 \tag{7}$$

where I^j represents the j^{th} frame of the predicted facial animation video and I_{GT} is the ground-truth.

3.4 Lip Sync Module

Because the other loss items guide the model to learn a averaged facial expression, so that the facial animation result will encounter a so-called "mean face" problem. Specifically, the generated or animated faces will have a tendency to converge towards an neutral expression, resulting in poor lip synchronization performance. To solve this problem, we introduce a lip sync loss item to make the movements of the lips be properly associated with the driving speech audio.

Because the lip motion is highly related to the driving audio, we can get a heuristic method that directly control the opening and closure of the mouth by analyzing the audio. Specifically, during the brief period before producing consonants like "b," "p," "m," and "f', the lips close together, while when producing vowels like "a," "o," and "i", the lips open wide.

To make use of this prior knowledge, we introduce a synchronization module to enhance the lip sync performance in the facial animation task. In detail, given the audio feature F_a extracted with HuBERT from the driving speech audio, we introduce a contrastive learning module that can control the lip motion to be more accurate and realistic.

In detail, as shown in Figure 3. given a short clip of talking head video, the corresponding audio, and a irrelevant audio clip, we use a lip expression encoder to get the lip movement feature F_{lip} and a audio feature encoder to get the audio-lip feature F_{al}. In the contrastive learning process, the lip expression of the video is set as the anchor sample, and the corresponding audio clip is set as the positive sample while the irrelevant one is set as the negative sample. The audio-lip feature of the positive and the negative sample are denoted as F^+_{al} and F^-_{al} respectively. We optimize the encoder in order to maximize the distance between F^-_{al} and F_{lip}, and minimize the distance between F^+_{al} and F_{lip}. During training, for each anchor sample, we select one positive sample and N negative samples ($N = 10$). To achieve this goal, we employ the Info noice-contrastive estimation (Info NCE) following CPC[41] as the loss function:

$$L_{lips} = -\log \frac{\exp(d(F_{lip}, F^+_{al})/\tau)}{\exp(d(F_{lip}, F^+_{al})/\tau) + \sum_N \exp(d(F, F^-_{al})/\tau)} \tag{8}$$

where $d(*)$ is a cosine distance function for 2 vectors:

$$d(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|_2 \|\mathbf{b}\|_2} \tag{9}$$

and τ is a temperature hyper-parameter set as 0.5.

4 Experiment

4.1 Dataset

The VOCA (Voice Operated Character Animation) dataset[35] is a large-scale audiovisual dataset specifically designed for training and evaluating models in the field of speech-driven facial animation. It aims to facilitate research and advancements in areas such as speech synthesis, speech recognition, and computer graphics.

The VOCAset comprises synchronized 4D facial scans, audio recordings, and 3D facial landmarks. It encompasses a diverse set of subjects, recording various facial expressions and speaking styles. The dataset includes 109 individuals covering diverse age groups, genders, and ethnic backgrounds.

Notably, the facial scans in the VOCA dataset capture not only the facial geometry but also the temporal

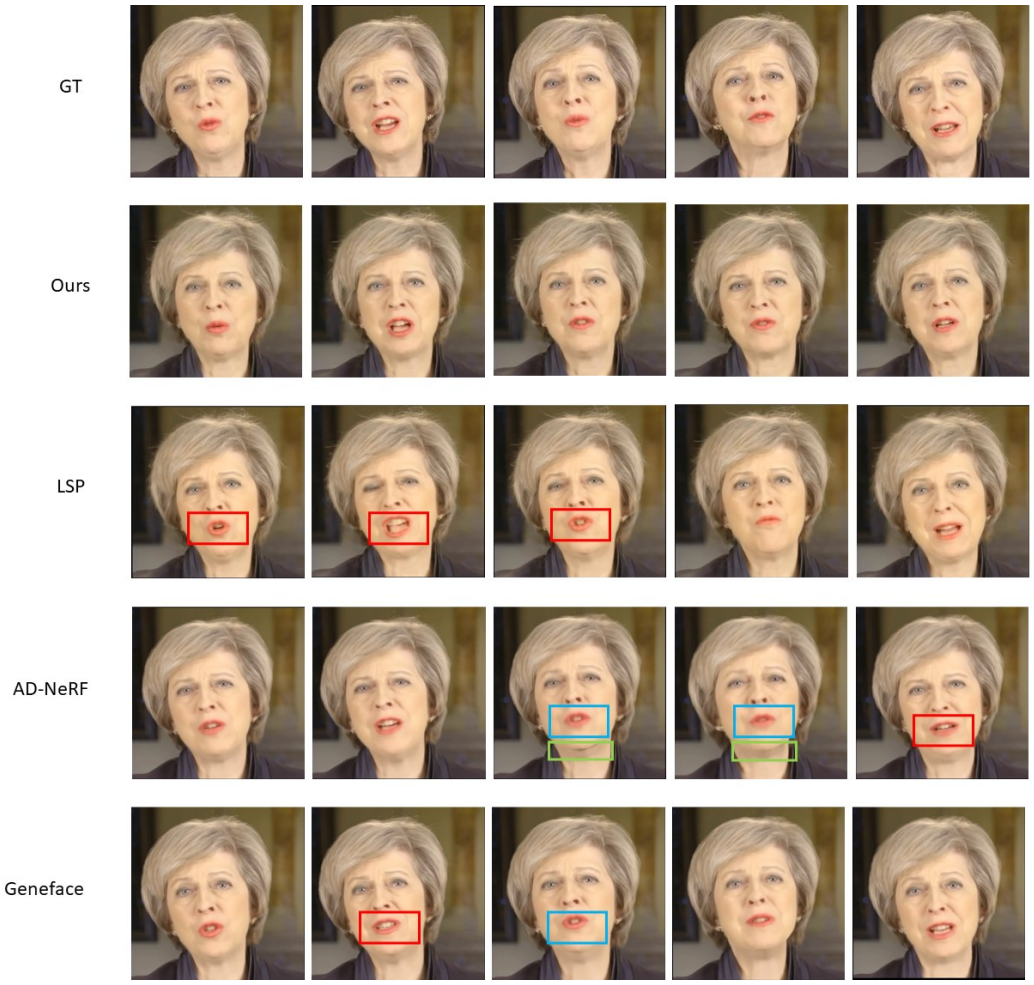

Fig.4. A qualitative result compared to other works including LSP[34], AD-NeRF[7] and Geneface[1]. The red box indicates the area where lip movements are inaccurate, the green box encloses the disharmony between the head and the body, and the blue box marks the blurry areas.

dynamics of facial movements. This temporal information is derived from high-quality RGB-D sequences and enables the generation of realistic and detailed facial animations synchronized with speech.

4.2 Implementation Details

We train our model with VOCAset. The Training process uses one NVIDIA RTX 3090 GPU. We carry out our experiments under the environment of python 3.7 with torch 1.10 and cuda-toolkit 11.1. We train our NeRF model for 800k iterations which takes about 60 hours.

We compare the performance of our model with several latest works including AD-NeRF[7] Geneface[1] Wav2lip[22].

4.3 Evaluation Metrics

To evaluate image quality, we make use of the FID score[42], PSNR, and SSIM as evaluation metrics. For evaluating audio-lip synchronization, we employ landmark distance (LMD)[43] and Sync-net confidence score (Sync Score)[22].

Frechet Inception Distance (FID) score measures the similarity between real and generated facial animations based on their feature distributions. It utilizes the Inceptionv3 neural network to extract features from

	PSNR	SSIM	FID	Sync score	LMD
Wav2lip	29.03	0.849	70.938	**7.329**	3.875
LSP	29.37	0.917	33.961	5.039	4.268
AD-NeRF	29.97	0.919	33.281	4.579	4.266
Geneface	30.14	0.934	**29.286**	5.346	3.571
Ours	**30.16**	**0.937**	30.294	5.236	**3.559**

Table 1. The quantitative comparison with different works. We use the videos proposed in [19] including Obama1, Obama2, May and Nadella as the testset. We use the released pre-trained model of other works to compare. Best results are in **bold**.

both real and generated frames. Lower FID scores indicate better similarity and higher quality in generated facial animations.

Structural Similarity Index (SSIM) measures the structural similarity between real and generated frames by evaluating luminance, contrast, and structural information. It assesses the overall quality and fidelity of generated facial animations. A higher SSIM score (ranging from 0 to 1) indicates better similarity and higher visual quality.

Peak Signal-to-Noise Ratio (PSNR) measures the quality of generated frames by comparing them to the original, real frames. It calculates the peak signal-to-noise ratio, which represents the ratio of the maximum possible power of a signal to the power of its noise. Higher PSNR values indicate higher fidelity and better quality in the generated facial animations.

Landmark Distance (LMD) measures the discrepancy between the landmarks (specific facial keypoints) in real and generated frames. It quantifies the accuracy and precision of facial alignment in the generated animations. Smaller landmark distances indicate better alignment and higher quality in the facial animation outputs.

SyncNet confidence score measures the alignment between the audio and visual aspects of a facial animation by analyzing the lip movements and corresponding audio signals with the publicly available pre-trained SyncNet[44]. The score is the average confidence score, higher scores indicate better audio-video correlation. See more detail in paper[44].

4.4 Quantitative evaluation

To validate our model, we conduct the evaluation experiment compared to several remarkable works. The quantitative results are reported in Table 1. It can be observed that our model achieves the best performance on most of the metrics on the test dataset.

On the PSNR, SSIM and LMD metrics, our model achieve the best compared with other baselines, indicating that the Disentangled NeRF has a great ability to render a high-fidelity photo of the human face with similar facial landmarks compared to the ground-truth. On the FID metric, although Geneface[1] model achieves the highest score, our model achieves a comparable result. As for the sync score metric, the proposed 2D method Wav2lip[22] achieves the best, but we also achieve a comparable performance with the best 3D-based method Geneface.

	PSNR	SSIM	FID	Sync score	LMD
Ours	**30.16**	**0.937**	**30.294**	**5.236**	**3.559**
Ours w/o sync module	29.58	0.916	34.562	4.593	4.245
Ours w/o disentangling	29.31	0.904	35.732	4.682	3.958

Table 2. The ablation study of the lip sync module and disentangling. We compare the performance between our model, our model without the lip sync module and our model without disentangling loss. Best results are in **bold**.

Fig.5. an example of qualitative result compared to AD-NeRF[7] and the ground truth. The lip movement synchronization of ours is much better than that of the baseline method.

4.5 Qualitative evaluation

As shown in Figure 4, our model can generate more realistic and natural result compared to the baselines. AD-NeRF uses 2 NeRF to render head and torso separately, which sometimes leads to distortion at the neck, which are marked in the green box in Figure 4. Besides, without a disentangling process, the lip area of the talking video generated by other methods appears to be blurred in some frames, which are marked in the blue box. Compared to the ground truth, the lip motion generated by all baseline models are different and inaccurate, which are marked in the red box. However, in out work, by utilizing a whole NeRF with disentangling loss and a sync module, we appropriately solve the problems and achieve a better performance.

Another example of qualitative comparison is shown in Figure 5. By comparing Ours with AD-NeRF[7] and the ground truth, Our results are far closer to the ground truth and have better temporal coherence and visual quality. A more intuitive comparison can be seen in the supplementary demonstration video.

As a 3D NeRF based facial method, our method is able to generate novel views from different camera poses. Figure 6 shows the pose manipulation results of our method. The results indicate that we can freely adjust head poses of the generated talking head within a range of perspectives, which is valuable in various applications.

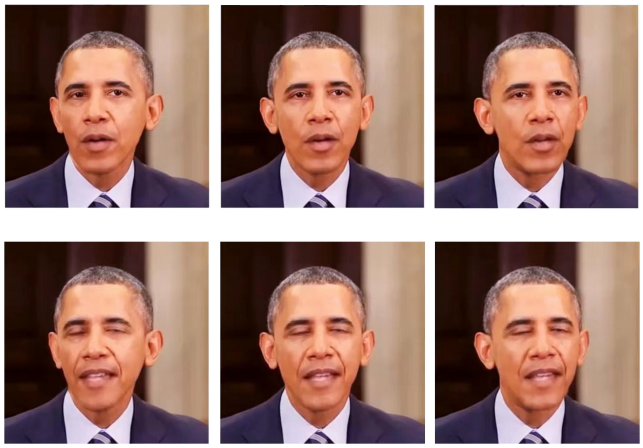

Fig.6. an example of head pose manipulation. Each row from left to right: novel view from left viewing direction, original view from middle, and novel view from right viewing direction.

4.6 Ablation Study

To validate the effectiveness of the proposed modules in the paper, we conduct several ablation studies. By systematically removing specific elements and components of our method, we can analyze the impact of these modules on the overall performance, thus knowing the significance of the modules.

Lip Sync Module To validate the effectiveness of the Lip Sync Module (Sec 3.4), we remove the lip sync module and retrain our NeRF without the lip feature F_{lip} as input. In this way, the audio feature can not directly contributes to the volume rendering process, thus resulting in a poor performance on the audio-video synchronization, especially in the areas around the lip. As shown in Table 2, the sync metrics including Sync score of the model without the lip sync module are significantly lower than the full model. Plus, the Landmark Distance of it is much higher than the full model, indicating the facial expression is significantly inaccurate compared to the model with the lip sync module.

Disentangling Loss To validate the effectiveness of the Disentangling loss (Sec 3.3), we remove the disentangling loss item and retrain our model. In this way, the learnable latent codes C will no longer be supervised to be uncoupled with one another. That is to say, the change in a person's expression will affect not only the latent code C_{exp}, but other latent codes as well. During the facial animation process, the latent codes all change irregularly, thus resulting in an unnatural and unrealistic volume rendering result. As shown in Table 2, the metrics evaluating the quality of the generated frames including SSIM, PSNR of the model without the disentangling loss are significantly lower than the full model, indicating that the facial animation performance is worse compared to the model with the lip sync module.

5 Conclusion

In this paper, we have proposed a Disentangling Facial Neural Radiance Field for talking head synthesis, which completely disentangles the Identity, expression, albedo and illumination of the face, so that the movement of the lips and the expressions driven by the speech audio will be accurate and realistic. We argue that by using proper loss items and synchronization module, our model can generate natural and accurate talking face in an end-to-end manner. Our study shows that the proposed method achieves the superior performance generating an audio-driven talking head video.

References

[1] Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*, 2023.

[2] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European Conference on Computer Vision*, pages 666–682. Springer, 2022.

[3] 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*, pages 296–301. Ieee, 2009.

[4] V Blanz and T Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194. ACM Press, 1999.

[5] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[6] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[7] 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.

[8] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[9] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.

[10] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[11] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021.

[12] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022.

[13] ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2023.

[14] Amit Raj, Michael Zollhofer, Tomas Simon, Jason Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11733–11742, 2021.

[15] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2021.

[16] 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.

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

[18] 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 Transactions on Multimedia*, 2023.

[19] 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.

[20] 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.

[21] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

[22] 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.

[23] 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.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[26] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[27] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[28] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[29] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] Yuanxun Lu, Jinxiang Chai, and Xun Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021.

[35] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019.

[36] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18780, 2022.

[37] 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.

[38] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[39] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[40] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution.

In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016.

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[42] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[43] 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.

[44] 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.