# Emotional Semantic Neural Radiance Fields for Audio-Driven Talking Head

Haodong Lin[1], Zhonghao Wu[1], Zhenyu Zhang[2], Chao Ma[1](✉), and Xiaokang Yang[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
{icegreen,ufouso,chaoma,xkyang}@sjtu.edu.cn
[2] Youtu Lab, Tencent, China

**Abstract.** Generating audio-driven talking head videos is a challenging problem which receives considerable attention recently. However, the emotional expressions of the speaker are often ignored, although the emotion information is expressed in the audio signal. In this paper, we propose Emotional Semantic Neural Radiance Fields (ES-NeRF), an audio-driven method for generating high-quality and emotional talking head videos based on neural radiance fields. Our method extracts the content features and the emotion features of the audio as additional inputs to construct a dynamic neural radiance field, applies the semantic segmentation map to constrain the speaker's expression, generates a dynamic three-dimensional emotional facial semantic representation, and then synthesizes the final high-quality video through the semantic translation network. Experiments show that our method can achieve high-quality results with corresponding expressions for audios containing different emotions that surpass the quality of state-of-the-art talking head methods.

**Keywords:** Emotional talking head · Nerual radiance field · Semantic segmentation

## 1 Introduction

Synthesizing audio-driven high-quality talking head videos is a challenging problem and necessary in many practical applications, such as film-making [17], virtual video conferences and digital humans [40]. Recently, many methods have been proposed to generate high-quality talking heads, and the mouth shape is kept in synchronizing with the audio. At present, the advanced methods are divided into two categories: the methods based on generative adversarial networks (GAN) [4,14,33,35,37,39] and the methods based on neural radiance fields (NeRF) [10,13].

This work was supported by NSFC (61906119, U19B2035), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and CCF-Tencent Open Research Fund.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 532–544, 2022.
https://doi.org/10.1007/978-3-031-20500-2_44

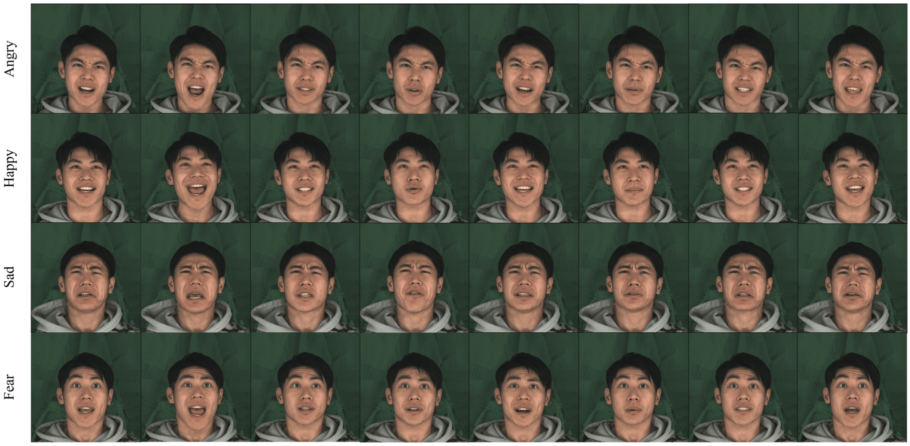

**Fig. 1.** Given an audio input, our ES-NeRF approach can generate high-quality emotional talking head videos. The example shows the generated talking heads with the same speech content but different emotions and poses.

Since it is difficult to directly learn the mapping from the original audio signal to the facial expression and the mouth shape, existing methods often use intermediate representations such as explicit 3D face shapes [26], 2D landmarks [31] or expression coefficients [33]. Nonetheless, there are very few works that consider the influence of the emotional factors contained in the audio signals on the expressions of the characters. The recently collected MEAD dataset [35] contains high-quality talking head videos with annotations of both emotion category and intensity.

NeRF-based methods have shown excellent results in recent years. For example, both AD-NeRF [13] and 4D facial avatar [10] can generate high-quality talking head videos with lip synchronization, and can ensure the free view direction and background switching. Nevertheless, AD-NeRF does not consider the influence of the speaker's emotions expressed on the face, and the rendering of some extreme expressions is heavily blurred, while the 4D facial avatar needs to be given the expression code of the target person and cannot be directly driven by audio signals.

To solve the problem current methods neglecting the influence of the speaker's emotion, we propose the Emotional Semantic Neural Radiance Fields (ES-NeRF), which can synthesize the high quality talking head videos with the correct expressions from emotional audio inputs, as the examples shown in Fig. 1. Different from existing NeRF-based methods, we consider to use both content features and emotional features of the input audio. For situations where it is difficult to directly learn the mapping from audio features to facial expressions and mouth shapes, we add semantic maps as additional supervisory signals. To solve the problem that dynamic neural radiance fields is difficult to handle the representation of multiple widely different expressions, we employ the generated three-dimensional semantic representation as an intermediate representation and produce the final result through a semantic translation network. The use of neural radiance fields to obtain three-dimensional semantic representations can naturally and freely adjust different postures and view directions, which is impossible in traditional 2D landmarks. For generating a single expression or multiple types of expressions, our model achieves better image quality, emotion accuracy and audio-visual synchronization than existing GAN-based or NeRF-based methods.

The main contributions of this work are listed as follow:

- We propose the Emotional Semantic Neural Radiance Fields, which is the first attempt to achieve emotional talking head generation in audio-driven NeRF-based methods.
- We add semantic segmentation constraints to the dynamic neural radiance fields to better learn the mapping from audio features to facial expressions and mouth shapes and meanwhile predict a 3D dynamic semantic representation.
- We propose the NeRF-to-GAN approach with adopting a semantic translation network to solve the problem of poor image quality directly rendered by dynamic neural radiance fields under multiple widely different expressions.

## 2 Related Work

### 2.1 Talking-Head Generation

**Generative Adversary Network.** As one pillar of the computer vision, GAN [11] contributes largely to this domain. Many researchers are inspired by this method and generate more and more realistic images and videos. Progressive GAN [15] can get high resolution results by applying the progressive training strategy on generator and discriminator. Style GAN [16] gets break through on the details and gets the ability to control the detail. Photo-realistic Audio-driven Video Portraits [37] and Neural voice puppetry [33] utilize a U-Net-based [27] GAN as a neural renderer to render 3DMM [2] parameters to realistic images and videos. Make-it-talk [39] disentangles the content feature and style feature from the audio feature, then choose landmarks as intermediary to generate final videos. Hierarchical Cross-Modal Talking Face Generation [4] proposes a cascade GAN with dynamic pixel-wise loss to avoid subtle artifacts and temporal discontinuities. These methods based on GAN all suffer the 3D inconsistency problem and neglect the emotional expression.

**Neural Radiance Field.** Since 2020, due to 3D consistency and remarkable neural rendering, Neural Radiance Field [21] has caught many eyes on it. Many methods [10,12,13,20,23,28,29] base on NeRF to set up 3D radiance fields and add additional parameters to control the generation of the results. Graf [29] combines NeRF and GAN framework to improve the quality of the image generation. Giraffe [23] extracts the feature map and then feeds it into GAN for progressive training to render high-resolution image. Style-NeRF [12] solves the problem in Giraffe that cannot generate high resolution image and produces the artifacts. AD-NeRF [13] adopts DeepSpeech [1] features of audios in NeRF to catch targets' lip movements and presents torso NeRF to render the torso part of the target. Similar to AD-NeRF, 4D facial avatars [10] chooses expression code estimated by 3DMM [2] model to match the lip movements and add the latent code to improve stability. Both AD-NeRF and 4D avatars maintain the talking head pose, while GNeRF [20] predicts head poses with the use of GAN and applies the generated pose in NeRF to render the talking head video. All aforementioned works cannot reenact the emotion styles of people.

### 2.2 Emotion Condition Generation

Many methods focus on generating a more high-resolution picture and a clearer mouth type, which lack consideration of personal style [33,37,39] and emotional style [8,14,22,25,35]. Previous works [10,33,37] generate videos with expressions by transferring the expression from source to target. These methods are limited to the source video and cannot generate emotion videos from unrecognized audios. Emotional style methods are inspired by voice emotionally recognizing methods [7,9,19] and disentangle the emotion in the audio by the depth network to generate emotional talking head. ExprGAN [8] designs one encoder-decoder architecture to control the expression identity with an expression controller module. Wang et al. [35] propose the MEAD dataset which contains high resolution talking head videos with eight different emotions and propose a network to generate the audio-driven lower face and emotion-driven upper face. GANimation [25] proposes one conditional GAN based on Action Units to estimate continuous facial movements of a designated expression. Audio EVP [14] disentangles content encoding and emotion encoding from the audio signal, then uses landmarks as intermediary to get the edge map. Finally EVP utilizes conditional GAN to translate the edge map to high-fidelity emotion video. Our work feeds content parameters and emotion parameters into NeRF to generate semantic results with expressions. Taking full advantage of both high-resolution generated image of GAN and 3D continuity of NeRF, our method can synthesize realistic high-quality videos with arbitrary head poses and free view directions. To this end, we employ semantic results as intermediate content.

## 3 Method

### 3.1 Overview

The framework of our Emotional Semantic Neural Radiance Fields is shown in Fig. 2. In order to obtain the meaningful information in expression from acoustic signals, we extract content features and emotion features separately. Inspired by 4D facial avatars [10], we use conditional implicit function and volume rendering to model the emotional dynamic talking head (Sec. 3.2). To better learn the mapping from audio features to facial expressions and mouth shapes, we make the dynamic neural radiance field understand the semantics of the human head

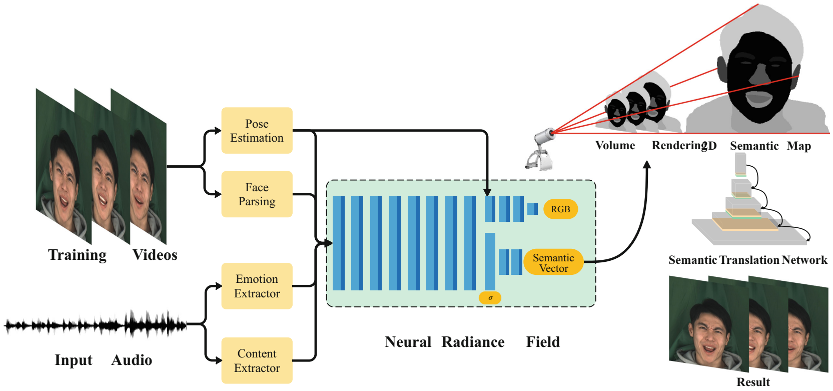

**Fig. 2.** Overview of our emotional semantic neural radiance fields Algorithm. We first extract content features and emotion features from the audio signal. Then we use the dynamic neural radiance field to predict 3D semantic representations. Finally, semantic translation network is applied to generate the high-quality rendering results with the corresponding emotion based on the semantic representations from the specific view direction.

and meanwhile predict its 3D semantic representation (Sec. 3.3). The NeRF-to-GAN module is designed to restore some facial details so that the model has better performance under the multiple types of emotions (Sec. 3.4). Besides, we focus on the facial emotional expression to generate natural emotional and speaking styles of the target person. Since the torso part has little relevance to the emotional expression of the characters, we process the head and torso part separately by segmenting the head and torso part and extracting a pure background, similar to AD-NeRF [13].

### 3.2 Emotional Neural Radiance Fields

Based on the dynamic neural rendering idea and to solve the problem of the neglect of emotion in neural radiance fields for talking heads, we employ the emotion feature and the content feature from the audio signal as additional inputs of the implicit function. In other words, spatial coordinates $\mathbf{x} = (x, y, z)$, viewing direction $\mathbf{d} = (\theta, \phi)$, content feature $\mathbf{m}$, and emotion feature e are meanwhile inputted to the implicit function which is realized by multi-layer perceptrons (MLPs) to implicitly represent the continuous 3D scene density $\sigma$ and color $\mathbf{c} = (r, g, b)$. The density $\sigma$ is the differential probability of a ray terminating at an infinitesimal particle at spatial coordinates $x$, which is only related to 3D position. The color $\mathbf{c}$ is RGB values, which can be predicted as a function of both spatial coordinates $\mathbf{x}$ and viewing direction $\mathbf{d}$. The entire implicit function can be formulated as follows:

$$F_\theta : (\mathbf{x}, \mathbf{d}, \mathbf{m}, \mathbf{e}) \rightarrow (\sigma, \mathbf{c}) \tag{1}$$

Like NeRF [21], we apply the volume rendering by numerical quadrature with hierarchical stratified sampling to compute the color of each pixel. Within one hierarchy, we mark a camera ray emitted from the center of projection of camera space through a given pixel as $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$ is the origin of the ray, $d$ is the view direction, and the near and far boundaries of $t$ are $t_n$ and $t_f$ respectively. Then the color of this ray can be expressed as an integral (numerically estimated by quadrature):

$$C(\mathbf{r}; p, \mathbf{m}, \mathbf{e}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\, dt \tag{2}$$

where $T(t)$ denotes the accumulated transmittance along the ray from $t_n$ to $t$. $p$ is the estimated head pose for transforming the sampling points to the canonical space like 4D facial avatars [10].

$$T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds) \tag{3}$$

To extract the content feature and emotion feature, we apply the popular DeepSpeech [1] model and OpenSmile [9] model respectively to obtain the DeepSpeech features and OpenSmile features. The 29-dimensional DeepSpeech features of 16 continuous frames are then sent into a temporal convolutional network to the temporally filtered content feature where we employ the self-attention idea [33]. The high-dimensional Opensmile features of each utterance are then normalized by Min-Max and feature selection based on L2 normalization to reduce the feature size to 100 dimension [30]. The low-dimensional emotion feature and the 76-dimensional filtered content feature are used as the inputs of MLPs.

### 3.3 3D Semantic Representation for Talking Head

During the implementation of rendering the emotional talking head by neural radiance fields, we find that the edges of the face and facial features of the generated results often produce some serious artifacts due to the large or extreme expressions. Since semantic representation is highly correlated with geometry and radiance reconstruction [28], we consider that significant high quality semantic labelling information could feasibly improve reconstruction quality.

We assume that one 3D position $\mathbf{x}$ has one semantic attribute $\mathbf{s}$, which denotes a distribution on $n$ semantic categories and has no relation with view directions. We introduce it into the above-mentioned dynamic neural radiance field, and map the input spatial coordinates and audio features (content features and emotion features) to semantic representation $\mathbf{s}$ through an implicit function. The specific implementation is to pass the 256-dimensional feature vector obtained after 8 fully-connected layers in MLPs through two additional fully-connected layers and a softmax normalisation layer that outputs the view-independent semantic representation. The entire implicit function can be formulated as follows:

$$F_\theta : (\mathbf{x}, \mathbf{d}, \mathbf{m}, \mathbf{e}) \rightarrow (\sigma, \mathbf{c}, \mathbf{s}) \tag{4}$$

The semantic representation of the pixels projected on the image plane through the ray can be expressed as an integral (numerically estimated by quadrature):

$$S(\mathbf{r}; p, \mathbf{m}, \mathbf{e}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{s}(\mathbf{r}(t), \mathbf{d})dt \tag{5}$$

where $T(t)$ denotes the accumulated transmittance along the ray from $t_n$ to $t$.

$$T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds) \tag{6}$$

### 3.4 NeRF-to-GAN

Although our method has been able to achieve the best results for rendering single-category emotions or similar emotions, when it is applied to reconstruct multiple categories of extremely different emotions, the overall clarity of the rendering results will decrease, especially the mouth part. Because semantic presentation performing better for above situation, we introduce a semantic translation network to generate corresponding image results. Following [3], we adopt a conditional-GAN architecture for the semantic translation network. We perform ray tracing on semantic occupancy field to obtain a 2D segmentation map from

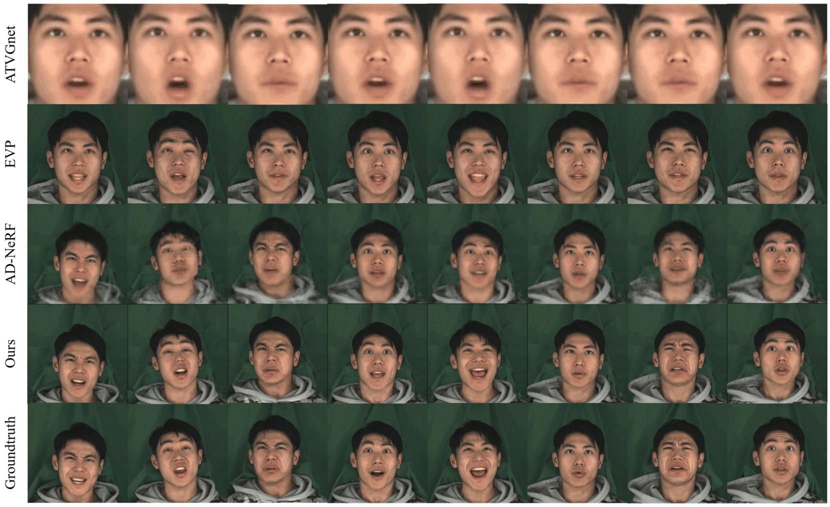

**Fig. 3.** Qualitative comparisons with the state-of-the-art methods. From up to down: Audio Transformation and Visual Generation Network (ATVGnet) [4], Audio-Driven Emotional Video Portraits (EVP) [14], Audio Driven Neural Radiance Fields (AD-NeRF) [13], Ours and the ground truth images. We show eight examples with different speech contents and eight different emotions. From left to right: angry, contempt, disgusted, fear, happy, neutral, sad and surprised.

a given user-specific viewpoint, then use the GAN generator to texture each semantic region from the style code sampled from the texture space. Finally we adopt a Semantic Instance Wise StyleGAN to regionally stylize the generated segmentation maps. Please refer to [3] for more details about the network architecture.

## 4 Experiments

### 4.1 Implementation Details

**Dataset.** We evaluate our method on the MEAD [35] dataset, a high-quality audio-visual dataset for emotional talking head generation with 60 actors and actresses and eight emotion categories. The dataset is split into the training and testing sets for models to train and test. Every emotional talking head video is converted to 25 fps and the sample rate of audio signals is set to be 16kHz.

**Data Preprocessing.** (1) Pose estimation. We firstly use an off-the-shelf method [38] to detect facial landmarks and align all the head part. The dynamics of the head pose are estimated by a state of-the-art face tracking approach [34] and bundle adjustment [32] approach is applied for optimization. The optimized head pose parameters are used for transforming the sampling points on the rays from the head part to the canonical space. (2) Face parsing. We use the popular face parsing method [5] to label the different semantic regions. Since we need to use the 2D segmentation result as the ground truth as supervisory signals for dynamic neural radiance fields, we cannot tolerate any obvious artifacts, such as the confusion between the left and right eyebrows. We reprocess the segmentation result by ignoring the distinction between the left and right parts to enhance the robustness to avoid the instability of the neural network training.

**Training Details.** We implement our framework in PyTorch [24]. Training images are resized to $512 \times 512$ for all the experiments. Neural radiance fields part are trained with Adam [18] solver with initial learning rate 5e-4. We train the neural network for 400k iterations and train the semantic translation network for 500k iterations. In each iteration of the neural network training, we randomly sample a batch of 2048 rays through the image pixels. For the target person, we choose 30 videos for each kind of emotions. We train both networks with a Tesla V100 with 32GB memory and need about 72 h with videos of single person with eight different emotions in resolution $512 \times 512$.

### 4.2 Evaluation Results

**Evaluation Metrics.** To quantitatively evaluate the expression accuracy, we extract facial landmarks from the generated videos and the ground truth videos. The metrics of Landmark Distance (LD) [4] and Landmark Velocity Difference (LVD) [39] are utilized to evaluate facial movements. LD represents the average Euclidean distance between ground truth and generated landmarks. LVD

**Table 1.** Quantitative comparisons with the state-of-the-art methods. We calculate the landmark accuracy, audio-visual consistency and video qualities of the results of different methods by comparing them with the ground truth images. ↑ indicates that the performance is better with higher results. ↓ indicates that the performance is better with smaller numbers.

| Method/Score | LD↓ | LVD↓ | SyncNet score↑ | PSNR↑ | SSIM↑ | Pose |
|---|---|---|---|---|---|---|
| ATVGnet [4] | 3.82 | 1.71 | 4.34 | 28.55 | 0.60 | static |
| EVP [14] | 3.01 | 1.56 | 5.17 | 29.53 | 0.71 | copied |
| AD-NeRF [13] | 3.42 | 1.71 | 4.56 | 28.20 | 0.82 | arbitrarily |
| **Ours** | **2.89** | **1.46** | **5.54** | **29.80** | **0.91** | **arbitrarily** |

denotes the average velocity differences of landmark movements between two videos. SyncNet [6] is often used to evaluate the audio-visual consistency for lip synchronization and facial motions. To evaluate the audio-visual synchronization quality, we use a pre-trained SyncNet model to compute the audio-sync offset and confidence of audio-driven talking head videos. The performance is better with higher score. Besides, we use PSNR and SSIM [36] to evaluate the quality of the generated images.

**Comparison with GAN-Based Methods.** We mainly choose the EVP [14] which has the best performance for emotional talking head generations in GAN-based methods to compare with our method. From the quantitative comparison in Table 1, we can see that our method performs better. In addition, the EVP method requires simultaneous input of audio signals and the source video to generate the talking head with poses, and its head postures are copied from the source video. In contrast, the posture and view direction of the talking head generated by our method can be adjusted arbitrarily, which is due to the ability of the neural radiance field to generate semantic representation with a free perspective. Analyzing and comparing the rendering results, we can find that the rendering results of EVP cannot well reflect the expression in some emotions (such as angry and disgusted).

**Comparison with NeRF-Based Methods.** AD-NeRF [13] is currently one of the few audio-driven NeRF-based talking head methods, and the input conditions are the same as our method: only audio input is required. It can be seen from the quantitative comparison in Table 1 that our method catches much higher score than AD-NeRF. It can be seen from the examples in Fig. 3 that although AD-NeRF can also express the speaker's emotion to a certain extent, its robustness is very poor and the rendering results are quite blurry, especially in the mouth area.

### 4.3 Ablation Study

To synthesizing high quality results, we add the emotional-related part, semantic-related part and semantic translation network into the dynamic NeRF. We select

LD, SyncNet score and PSNR as metrics for the ablation study to demonstrate the necessity of these parts in our method.

**Table 2.** Quantitative ablation study. Emo. denotes the emotion-related part of dynamic neural radiance fields. Sem. denotes the semantic-related part of dynamic neural radiance fields. Trans. denotes the semantic translation network

| Emo. | Sem. | Trans. | LD↓ | SyncNet score ↑ | PSNR ↑ |
|---|---|---|---|---|---|
| | ✓ | | 3.36 | 4.75 | 28.5 |
| ✓ | | | 3.17 | 5.02 | 29.4 |
| ✓ | ✓ | | 3.05 | 5.23 | **30.2** |
| ✓ | ✓ | ✓ | **2.89** | **5.54** | 29.8 |

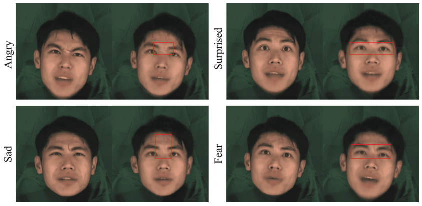

**Fig. 4.** Ablation study for emotional neural radiance fields. We show cases with (left) and without (right) emotion related part in neural radiance field. The red boxes show the artifacts in the generated frame. (Color figure online)

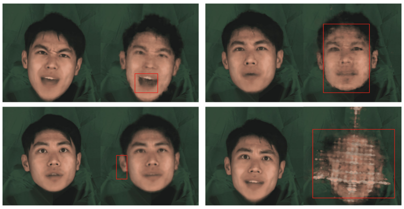

**Fig. 5.** Ablation study for 3D semantic representation. We show cases with (left) and without (right) semantic related part in neural radiance field. The red boxes show the artifacts in the generated frame. (Color figure online)

**Emotional Neural Radiance Fields.** The emotion feature as an additional input of dynamic NeRF is helpful for rendering the talking head with real emotion. In Table 2, without emotion related part in neural radiance field, all evaluation metrics perform worse. Figure 4 shows that lack of the constraints of emotional features, synthesized results are less robust in emotional representation, and often present incorrect expressions.

**3D Semantic Representation for Talking Head.** The semantic-related part can not only be used to correctly predict the semantic representation of the head, but also facilitate the image rendering of the NeRF. In Table 2, the result of adding the semantic part improves on all evaluation metrics. It can be seen from the Fig. 5 that the semantic part well constrains the image rendering results on the edge of each organ, making it smoother and clearer.

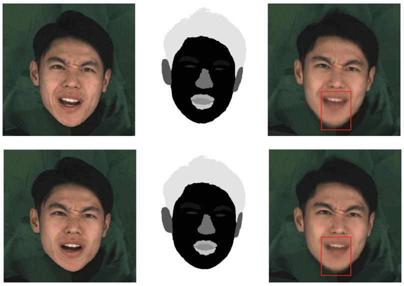

**Fig. 6.** Ablation study for semantic translation network. We show the cases with (left) and without (right) semantic translation network and predicted semantic representation (medium). The red box shows the artifacts in the generated frame. (Color figure online)

**Semantic Translation Network.** It is worthy to mention that it is difficult to generate high-quality 3D semantic representations without the emotional part or semantic part in the dynamic NeRF. Under these situations, the result with the semantic translation network is worse than directly rendered by the dynamic NeRF. Therefore, the semantic translation network is removed in the subsequent comparisons where the emotional part or semantic part in the dynamic NeRF is removed. In Table 2, both LD and SyncNet score perform worse, but PSNR improves without the semantic translation network. However, we can see from the rendered sample image in Fig. 6 that the overall clarity of the result is actually lower, and the image quality is not as good as the result of the complete method. Especially in the mouth area, it is difficult to recognize the mouth shape of the speaker. Therefore, the final solution adds the semantic translation network.

## 5 Conclusion

In this paper, we propose an audio-driven talking head method for generating high-quality and emotional portraits videos based on neural radiance fields. Our proposed emotional neural radiance field uses the content features and the emotion features extracted from the audio as inputs to reconstruct the emotional talking head. Then, we employ semantic segmentation to constrain the speaker's expression and generate the three-dimensional dynamic facial semantic representation. To improve the quality of synthesized results, we propose a NeRF-to-GAN approach and generate the final high-quality video containing different emotions through the semantic translation network. The photo-realistic generated results surpass the quality of state-of-the-art talking head methods both quantitatively and qualitatively.

## References

1. Amodei, D., et al.: Deep speech 2: end-to-end speech recognition in English and Mandarin. In: ICML (2016)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH (1999)
3. Chen, A., Liu, R., Xie, L., Chen, Z., Su, H., Jingyi, Y.: Sofgan: a portrait image generator with dynamic styling. TOG **41**(1), 1–26 (2021)
4. Chen, L., Maddox, R.K., Duan, Z., Xu, C.: Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: CVPR (2019)

5. Cheng-Han Lee, Ziwei Liu, L.W., Luo, P.: Maskgan: towards diverse and interactive facial image manipulation. In: CVPR (2020)
6. Chung, J.S., Zisserman, A.: Out of time: automated lip sync in the wild. In: Chen, C.-S., Lu, J., Ma, K.-K. (eds.) ACCV 2016. LNCS, vol. 10117, pp. 251–263. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54427-4_19
7. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Sig. Process. Mag. **18**(1), 32–80 (2001)
8. Ding, H., Sricharan, K., Chellappa, R.: Exprgan: facial expression editing with controllable expression intensity. In: AAAI (2018)
9. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: ACMMM (2010)
10. Gafni, G., Thies, J., Zollhofer, M., Niessner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: CVPR (2021)
11. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS (2014)
12. Gu, J., Liu, L., Wang, P., Theobalt, C.: Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. arXiv preprint. arXiv:2110.08985 (2021)
13. Guo, Y., Chen, K., Liang, S., Liu, Y.J., Bao, H., Zhang, J.: Ad-nerf: audio driven neural radiance fields for talking head synthesis. In: ICCV (2021)
14. Ji, X., et al.: Audio-driven emotional video portraits. In: CVPR (2021)
15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability and variation. In: ICLR (2018)
16. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
17. Kim, H., et al.: Neural style-preserving visual dubbing. TOG **38**(6), 1–13 (2019)
18. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
19. Kwon, O.W., Chan, K., Hao, J., Lee, T.W.: Emotion recognition by speech signals. In: EUROSPEECH (2003)
20. Meng, Q., et al.: Gnerf: gan-based neural radiance field without posed camera. In: ICCV (2021)
21. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 405–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24
22. Mittal, G., Wang, B.: Animating face using disentangled audio representations. In: WACV (2020)
23. Niemeyer, M., Geiger, A.: Giraffe: representing scenes as compositional generative neural feature fields. In: CVPR (2021)
24. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: NIPS (2019)
25. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: anatomically-aware facial animation from a single image. In: ECCV (2018)
26. Ran, Y., Zipeng, Y., Juyong, Z., Hujun, B., Yong-Jin, L.: Audio-driven talking face video generation with natural head pose. In: ICCV (2021)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. S. Zhi, T. Laidlow, S.L., Daviso, A.J.: In-place scene labelling and understanding with implicit scene representation. In: ICCV (2021)

29. Schwarz, K., Liao, Y., Niemeyer, M., Geiger, A.: Graf: generative radiance fields for 3d-aware image synthesis. In: NIPS (2020)
30. Sebastian, J., Pierucci, P., et al.: Fusion techniques for utterance-level emotion recognition combining speech and transcripts. In: Interspeech (2019)
31. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. TOG **36**(4), 1–13 (2017)
32. T. Baltrusaitis, M.M., Robinson, P.: Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: FG (2015)
33. Thies, J., Elgharib, M., Tewari, A., Theobalt, C., Nießner, M.: Neural voice puppetry: audio-driven facial reenactment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12361, pp. 716–731. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58517-4_42
34. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: CVPR (2016)
35. Wang, K., et al.: MEAD: a large-scale audio-visual dataset for emotional talking-face generation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12366, pp. 700–717. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58589-1_42
36. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP **13**(4), 600–612 (2004)
37. Wen, X., Wang, M., Richardt, C., Chen, Z.Y., Hu, S.M.: Photorealistic audio-driven video portraits. TVCG **26**(12), 3457–3466 (2020)
38. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: a boundary-aware face alignment algorithm. In: CVPR (2018)
39. Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., Li, D.: Makeittalk: speaker-aware talking-head animation. TOG **39**(6), 1–15 (2020)
40. Zhou, Y., Xu, Z., Landreth, C., Kalogerakis, E., Maji, S., Singh, K.: Visemenet: audio-driven animator-centric speech animation. TOG **37**(4), 1–10 (2018)