Supplementary Material
Parametric Implicit Face Representation for Audio-Driven Facial Reenactment

Ricong Huang [1] Peiwen Lai[1] Yipeng Qin[2] Guanbin Li[1*]
[1]School of Computer Science and Engineering, Sun Yat-sen University [2]Cardiff University
{huangrc3, laipw5}@mail2.sysu.edu.cn, qiny16@cardiff.ac.uk, liguanbin@mail.sysu.edu.cn

In this supplement, we provide more implementation details of the network architecture of the three components we proposed: contextual audio to expression encoding (Sec. 1), implicit representation parameterization (Sec. 2), rendering with parametric implicit representation (Sec. 3). Please note that we also include an additional ablation study on the choice of hyper-parameter k in Sec. 1. Lastly, we show additional qualitative evaluation results in Sec. 4. We strongly encourage readers to watch our supplementary video, which demonstrates the superiority of our method.

1. Contextual Audio to Expression Encoding

1.1. Network Architecture

As Fig. 1 shows, our contextual audio to expression encoding component is a transformer-based architecture similar to [4], consisting of a transformer encoder and decoder.

For the transformer encoder, we first extract the primary audio feature of a raw audio A through wav2vec 2.0 [1], which consists of an audio feature extractor and a multi-layer transformer encoder. Between them, the audio feature extractor consists of several temporal convolutions layers (TCN), and the transformer encoder is a stack of multi-head self-attention and feed-forward layers. Note that we have added a linear interpolation layer in-between to resample the audio features from the TCN output to ensure that they share the same sampling frequency with the training video. The dimension of each block is 1024 and the number of attention heads is 16. A linear projection layer is added after the transformer blocks to project the extracted features to the input space of the biased cross-modal MH Attention blocks in the transformer decoder.

The transformer decoder takes input from the output of the transformer encoder, the style embedding layer, and the expression encoder. Its output is converted by the expression decoder into the predicted expression parameter $a_1, a_2, ..., a_k$. We use a sequence of $k = 100$ video frames for training. Among them, the style embedding and expression encoder are both fully-connected layers with a dimension of 1024; the expression decoder is a fully-connected layer with a dimension of 64. The style embedding layer takes the identity parameter z_{id} as input. The expression encoder takes the previous predicted expression parameter $a_1, a_2, ..., a_{k-1}$ as input. Their outputs are summed together except for the first frame when only z_{id} is available. Like [4], each block in the transformer decoder consists of a periodic positional encoding layer, a biased causal multi-head self-attention layer, a biased cross-modal multi-head attention layer and a feed forward layer, whose details are described as follows.

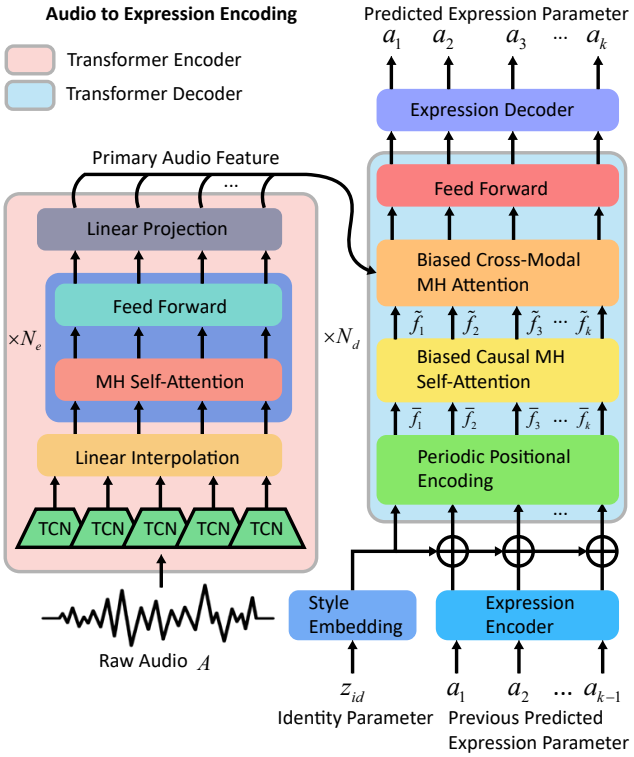

Figure 1. Network architecture of our contextual audio to expression encoding component.

*Corresponding author is Guanbin Li.

Periodic Positional Encoding (PPE). PPE is used for the injection of temporal order and is formulated as:

$$\begin{aligned} PPE_{(t,2i)} &= \sin\left((t \mod p)/10000^{2i/d}\right) \\ PPE_{(t,2i+1)} &= \cos\left((t \mod p)/10000^{2i/d}\right) \end{aligned}, \quad (1)$$

where $p = 25$ indices, the period t denotes the current time-step in the input sequence, d is its dimension and i is the dimension index. The output of PPE is a sequence of features $\overline{F}_t = (\overline{f}_1, \ldots, \overline{f}_t)$, $1 \leq t \leq k$.

Biased Causal Multi-head (MH) Self-attention. This layer is designed to ensure causality and to improve the generalization of the model to long sequences, which is formulated as:

$$\begin{aligned} &\mathrm{MH}(Q^{\overline{F}}, K^{\overline{F}}, V^{\overline{F}}, B^{\overline{F}}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_\mathrm{H}) W^{\overline{F}} \\ &\mathrm{head}_\mathrm{h} = \mathrm{softmax}\left(\frac{Q_h^{\overline{F}}(K_h^{\overline{F}})^T}{\sqrt{d_k}} + B_h^{\overline{F}}\right) V_h^{\overline{F}} \\ &B^{\overline{F}}(i,j) = \begin{cases} \lfloor (i-j)/p \rfloor, & j \leq i \\ -\infty, & \text{otherwise} \end{cases} \\ &B_h^{\overline{F}} = B^{\overline{F}} m \end{aligned}, \quad (2)$$

where $Q^{\overline{F}}, K^{\overline{F}}, V^{\overline{F}}$ are projected from the sequence $\overline{F}_t = (\overline{f}_1, \ldots, \overline{f}_t)$, $W^{\overline{F}}$ is a parameter matrix, d_k is the dimension of $Q^{\overline{F}}$ and $K^{\overline{F}}$, $B^{\overline{F}}$ is the temporal bias matrix and i, j are the indices of it, and m is a head-specific slope. The output of this layer is a sequence of features $\tilde{F}_t = (\tilde{f}_1, \ldots, \tilde{f}_t)$, $1 \leq t \leq k$.

Biased Cross-modal Multi-head (MH) Attention. This layer combines the output of the transformer encoder and $\tilde{F}_t = (\tilde{f}_1, \ldots, \tilde{f}_t)$, which is formulated as:

$$\begin{aligned} &\mathrm{Att}(Q^{\tilde{F}}, K^A, V^A, B^A) = \mathrm{softmax}\left(\frac{Q^{\tilde{F}}(K^A)^T}{\sqrt{d_k}} + B^A\right) V^A \\ &B^A(i,j) = \begin{cases} 0, & i \leq j < (i+1) \\ -\infty, & \text{otherwise} \end{cases} \end{aligned}, \quad (3)$$

where B^A is the alignment bias matrix. Equation (3) is extended to H heads to explore different subspaces.

Feed Forward Layer. A fully-connected layer of dimension 2048.

For both the biased causal MH self-attention and the biased cross-modal MH attention, 4 heads are employed with a model dimension of 1024. We use $N_e = 24$ and $N_d = 1$ transformer blocks in our implementation.

1.2. Ablation Study on Sequence Length k

As mentioned in our main paper, our audio to expression encoding is a stand-alone and light-weight task that can learn the contextual information from *long* audio sequences. To justify the benefit brought by long sequences, we conduct an ablation study on the sequence length k. As Tab. 1 shows, in general, the generated talking head videos achieve better LMD and AVConf scores with longer training audio sequences, indicating that capturing the long-term contextual information from audio sequences helps generate highly synchronized lip movements for talking portraits. We use $k = 100$ in our method as it achieves the best scores.

k	1	10	50	100	200
LMD↓	2.556	2.273	1.626	**1.477**	1.650
AVConf↑	4.201	4.505	6.873	**7.071**	6.929

Table 1. Ablation study of hyper-parameter k (sequence length). $k = 1$ indicates no use of contextual information.

2. Implicit Representation Parameterization

We implement our implicit representation parameterization based on an efficient tri-plane structure [2].

As shown in Fig. 2 of the main paper, we employ a mapping network to produce a 512-D intermediate latent vector z from the concatenation of identity parameter z_{id} and expression parameter z_{exp}. Conditioned on the latent vector z, a StyleGAN2 [6] generator is employed to generate a $96 \times 256 \times 256$ feature map. Then the feature map is split into three axis-aligned orthogonal feature planes, each with a resolution of $32 \times 256 \times 256$. Given camera pose R, t and intrinsic matrix K, a 3D point position in world coordinates $[x_w\ y_w\ z_w\ 1]^T$ is calculated based on:

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} R & t \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}, \quad (4)$$

where $[u\ v\ 1]^T$ is the 2D point position in pixel coordinates and z_c is the z-coordinate of 3D point position in camera coordinates. Then we project it onto each of the three feature planes, retrieve the feature vector $(F_{xy}, F_{xz}, F_{yz}$ via bilinear interpolation, and aggregate the three feature vectors via summation. These aggregated features are interpreted as 32-D color and 1-D density through a lightweight decoder which is a multi-layer perceptron with a single hidden layer and a softplus activation function. Furthermore, they are reconstructed as a $32 \times 64 \times 64$ feature map I_F using volume rendering. Two-pass importance sampling strategy is used to implement volume rendering [7] as in [8]. Compared to NeRF structures using large fully connected networks [8], the computational cost of tri-plane based neural rendering is reduced since it has a smaller decoder. Finally, a CNN-based upsampling network is used to upsam-

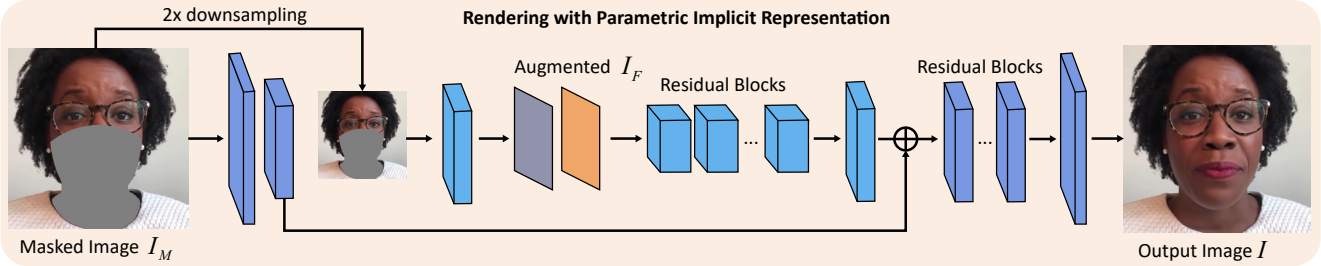

Figure 2. Network architecture of our rendering with parametric implicit representation component. We formulate facial reenactment as an image inpainting problem conditioned on the implicit representation I_F and use a coarse-to-fine network structure to tackle it.

Figure 3. Additional qualitative comparison with ATVG [3], Wav2Lip [9], MakeitTalk [13] and PC-AVS [12].

ple and render I_F to the final image I_{face} with a resolution of $3 \times 512 \times 512$.

3. Rendering with PIR

As shown in Fig. 2, our rendering with parametric implicit representation (PIR) component uses a coarse-to-fine generator [10] to generate the output image. The input $3 \times 512 \times 512$ masked image I_M is downsampled to a $3 \times 256 \times 256$ image through the average pooling. Then the image is further processed through several convolution layers and concatenated with the augmented feature map I_F from the implicit representation parameterization component. The concatenated features are further processed through 4 residual blocks and several convolution layers. After that, the features is summed with the feature maps extracted from I_M, passing through 3 residual blocks and finally rendering into the output image I through convolution layers.

4. Additional Qualitative Results

As a complement to the main paper, we show additional qualitative results of ATVG [3], Wav2Lip [9], MakeitTalk [13], PC-AVS [12] and our method. As Fig. 3 shows, it can be observed that our method generates talking portraits with

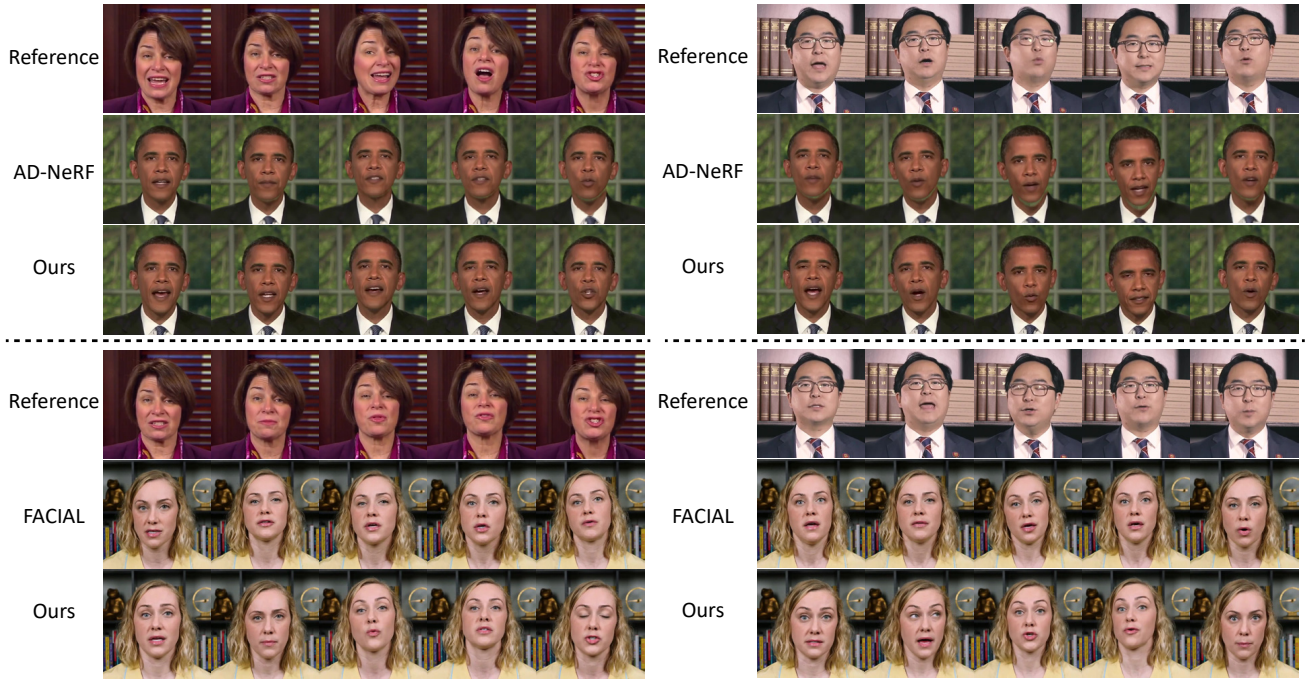

Figure 4. Comparison with AD-NeRF [5] and FACIAL [11].

highly synchronized lip movements and high fidelity to the facial details, outperforming all previous methods.

We also show more qualitative comparison results with AD-NeRF [5] and FACIAL [11] in Fig. 4. The generated talking portraits are driven by the audio from different identities. The results show that the talking heads generated by AD-NeRF [5] have obvious artifacts at the head-neck junction, and those generated by FACIAL [11] have less accurate lip movements. In contrast, our method can generate natural and vivid talking portraits, indicating that it generalizes better to unseen audios.

References

[1] 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. 1

[2] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[3] 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. 3

[4] 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. 1

[5] 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. 4

[6] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[7] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 2

[8] 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. 2

[9] 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. 3

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8798–8807, 2018. 3

[11] 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. 4

[12] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4176–4186, 2021. 3

[13] 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. 3