Parametric Implicit Face Representation for Audio-Driven Facial Reenactment

Ricong Huang 1 Peiwen Lai1 Yipeng Qin2 Guanbin Li1*
1School of Computer Science and Engineering, Sun Yat-sen University 2Cardiff University
{huangrc3, laipw5}@mail2.sysu.edu.cn, qiny16@cardiff.ac.uk, liguanbin@mail.sysu.edu.cn

Abstract

Audio-driven facial reenactment is a crucial technique that has a range of applications in film-making, virtual avatars and video conferences. Existing works either employ explicit intermediate face representations (e.g., 2D facial landmarks or 3D face models) or implicit ones (e.g., Neural Radiance Fields), thus suffering from the trade-offs between interpretability and expressive power, hence between controllability and quality of the results. In this work, we break these trade-offs with our novel parametric implicit face representation and propose a novel audio-driven facial reenactment framework that is both controllable and can generate high-quality talking heads. Specifically, our parametric implicit representation parameterizes the implicit representation with interpretable parameters of 3D face models, thereby taking the best of both explicit and implicit methods. In addition, we propose several new techniques to improve the three components of our framework, including i) incorporating contextual information into the audio-to-expression parameters encoding; ii) using conditional image synthesis to parameterize the implicit representation and implementing it with an innovative tri-plane structure for efficient learning; iii) formulating facial reenactment as a conditional image inpainting problem and proposing a novel data augmentation technique to improve model generalizability. Extensive experiments demonstrate that our method can generate more realistic results than previous methods with greater fidelity to the identities and talking styles of speakers.

1. Introduction

Audio-driven facial reenactment, also known as audio-driven talking head generation or synthesis, plays an important role in various applications, such as digital human, film-making and virtual video conference. It is a challenging cross-modal task from audio to visual face, which requires the generated talking heads to be photo-realistic and have lip movements synchronized with the input audio.

According to the intermediate face representations, existing facial reenactment methods can be roughly classified into two categories: explicit and implicit methods. Between them, explicit methods [5, 18, 27, 29, 30, 34, 37, 40, 44] exploit relatively sophisticated 2D (e.g., 2D facial landmarks [5, 18, 29, 34, 44]) or 3D (e.g., 3D Morphable Model [27, 30, 37, 40]) parametric face models to reconstruct 2D or 3D faces, and map them to photo-realistic faces with a rendering network such as the Generative Adversarial Networks (GANs) [32, 39]. Their distinct advantage is

*Corresponding author is Guanbin Li.
the controllability (e.g., expressions) resulting from their interpretable facial parameters. However, despite this advantage, the parametric face models used in explicit methods are often sparse and have very limited expressive power, which inevitably sacrifices the quality of synthesized faces (e.g., the inaccurate lip movements and blurry mouth caused by the missing teeth area in 3D face models). In contrast, implicit methods [12, 16, 17, 24, 25, 28, 42, 43] use implicit 2D or 3D representations that are more expressive and can generate more realistic faces. For example, Neural Radiance Fields (NeRF) based methods [12, 17, 25] are one of the more representative implicit methods that use NeRF to represent the 3D scenes of talking heads. Although being more expressive and producing higher-quality results, implicit methods are not interpretable and lose the controllability of the synthesis process, thus requiring model retraining to change its target person. As a result, the explicit and implicit methods mentioned above form a trade-off between the interpretability and expressive power of intermediate face representations, while a representation that is both interpretable and expressive remains an open problem.

In this work, we break the above trade-off by proposing a parametric implicit representation that is both interpretable and expressive, paving the way for controllable and high-quality audio-driven facial reenactment. Specifically, we propose to parameterize implicit face representations with the interpretable parameters of the 3D Morphable Model (3DMM) [10] using a conditional image synthesis paradigm. In our parametric implicit representation, the 3DMM parameters offer interpretability and the implicit representation offers strong expressive power, which take the best of both explicit and implicit methods (Fig. 1). To implement our idea, we propose a novel framework consisting of three components: i) contextual audio to expression (parameters) encoding; ii) implicit representation parameterization; iii) rendering with parametric implicit representation. Among them, our contextual audio to expression encoding component employs a transformer-based encoder architecture to capture the long-term context of an input audio, making the resulting talking heads more consistent and natural-looking; our implicit representation parameterization component uses a novel conditional image synthesis approach for the parameterization, and innovatively employs a tri-plane based generator offered by EG3D [3] to learn the implicit representation in a computationally efficient way; our rendering with parametric implicit representation component formulates face reenactment as an image inpainting problem conditioned on the parametric implicit representation to achieve a consistent and natural-looking “blending” of the head and torso of a target person. In addition, we observe that the model slightly overfits to the training data consisting of paired audio and video, causing jitters in the resulting talking heads whose lip movements are required to be synchronized with unseen input audio. To help our model generalize better and produce more stable results, we further propose a simple yet effective data augmentation strategy for our rendering component.

In summary, our main contributions include:

- We propose an innovative audio-driven facial reenactment framework based on our novel parametric implicit representation, which breaks the previous trade-off between interpretability and expressive power, paving the way for controllable and high-quality audio-driven facial reenactment.
- We propose several new techniques to improve the three components of our innovative framework, including: i) employing a transformer-based encoder architecture to incorporate contextual information into the audio to expression (parameters) encoding; ii) using a novel conditional image synthesis approach for the parameterization of implicit representation, which is implemented with an innovative tri-plane based generator [3] for efficient learning; iii) formulating facial reenactment as a conditional image inpainting problem for natural “blending” of head and torso, and proposing a simple yet effective data augmentation technique to improve model generalizability.
- Extensive experiments show that our method can generate high-fidelity talking head videos and outperforms state-of-the-art methods in both objective evaluations and user studies.

2. Related work

Given a video of a target person and an (unpaired) audio, audio-driven facial reenactment aims to synthesize a novel video of the target person whose lip movement is synchronized with the given audio. Most existing talking head generation methods can be roughly classified into two categories: explicit methods and implicit methods, according to their intermediate face representations.

Explicit Methods. Explicit methods use parametric face models as intermediate face representations. Depending on the type of parametric face models used, explicit methods can be further divided into two categories: 2D-based and 3D-based. Between them, 2D-based methods use 2D parametric face models like 2D facial landmarks [5, 18, 29, 34, 44], and map the input audio to them. These 2D landmarks are then fed into generative adversarial networks (GANs) to synthesize photo-realistic faces. For example, Chen et al. [5] propose an adjustable pixel-wise loss to guide the network to focus on audiovisual-correlated facial landmarks. Xie et al. [34] predict the facial landmarks in the mouth area with the input audio and then change the lip movement of
video frames to match the predicted landmarks. In contrast, 3D-based methods use more expressive 3D parametric face models (e.g., the 3D Morphable Models (3DMM) [2, 23] and FLAME [15]) and map the audio to the expression parameters of the models. These expression parameters, together with those extracted from the video frames, are used to reconstruct explicit 3D face shapes that will be fed into a rendering network to synthesize new talking head videos. For example, Thies et al. [30] encode the audio into a general audio-expression space and learn a person-specific expression basis to reconstruct the intermediate 3D model. Zhang et al. [40] leverage the context in audio to model implicit attributes like eye blinking and head pose, extending the attribute control of the face model. Song et al. [27] remove identity information in audio to improve the quality of expression parameters. And they exploit the use of expression and landmarks from video frames to supervise the reconstructed facial mesh. Despite being interpretable, all parametric face models used in explicit methods are sparse compared to image pixels and cannot capture facial details (e.g., the missing areas like teeth in 3D face models).

Implicit Methods. Some works achieve the audio-to-face transition directly through the Generative Adversarial Networks (GANs) [16, 24, 28, 36, 42, 43]. Prajwal et al. [24] employ a powerful lip-sync discriminator to detect lip-sync errors, forcing the generator to extract more expressive representations from the input audio. Zhou et al. [43] devise an implicit pose code to achieve free pose control and enhance the audio representation by contrastive learning in a non-identity space. Recently, some other implicit methods use Neural Radiance Fields (NeRF) [19] as the intermediate representation [12, 17, 25, 45], which models the 3D scene of a talking head with a fully-connected network and volume rendering techniques. For example, Guo et al. [12] employ two individual sets of NeRF to synthesize the talking head and torso of a portrait respectively. Liu et al. [17] leverage the semantic information in video frames to guide NeRF to concentrate on the hard-to-learn area like mouth and eyes. Shen et al. [25] introduce audio conditions to warp the face to the query space, which is applied in the fine-tuning of the facial radiance field for few-shot synthesis. Although implicit methods have more expressive representations and produce higher-quality videos, they are less interpretable and lose the controllability of the synthesis process, thus requiring model re-training to change its target person.

In this work, we propose a novel framework that takes the best of both explicit and implicit methods. Specifically, we exploit the interpretable parameter space of 3D parametric face models, but map them to implicit face representations instead of reconstructing 3D face models. In this way, we obtain a representation that is both expressive and interpretable, thus paving the way for controllable and high-quality audio-driven facial reenactment.

3. Methodology

As Fig. 2 shows, our framework consists of three components: contextual audio to expression (parameters) encoding (Sec. 3.1), implicit representation parameterization (Sec. 3.2) and rendering with parametric implicit representation (Sec. 3.3). Given an input raw audio \( A \), our contextual audio to expression encoding component maps \( A \) to its corresponding expression parameter \( z_{exp} \) in the same format as used in 3D Morphable Model (3DMM). \( z_{exp} \), together with the identity parameter \( z_{id} \) and the pose parameter (rotation \( R \), translation \( t \) and camera intrinsic matrix \( K \), which are used as the camera pose) extracted by 3DMM, constitute the facial parameters and are mapped to the implicit representation of a reenacted face \( I_F \) by our implicit representation parameterization component. Finally, our rendering with parametric implicit representation (PIR) component formulates facial reenactment as a conditional inpainting task and renders the reenacted image with \( I_F \) as the condition.

3.1. Contextual Audio to Expression Encoding

As Fig. 2 shows, unlike previous implicit methods that train the audio encoder in an end-to-end fashion [12, 17], we explicitly supervise its training with expression parameters extracted by 3D Morphable Model (3DMM). The rationale behind our choice is that audio has little to do with a person’s identity or pose but mainly his/her expression (e.g., lip movement). Specifically, given a raw audio \( A \), we first extract its preliminary feature using wav2vec 2.0 [1], a self-supervised pre-trained speech model that facilitates accurate lip movement through the abundant phoneme information it learned from a large-scale corpus of unlabeled speech. Then, we feed this feature along with the identity parameters extracted by 3DMM into a transformer-based audio feature extractor network [11] which encodes it to the expression parameters \( a_k \) of the \( k \)-th video frame. Thanks to the transformer architecture, \( a_k \) is dependent on the expression parameters of previous frames \( \{a_i | i = 1, 2, \ldots, k - 1\} \) and effectively captures the context of the audio. In addition, our method separates audio encoding as a stand-alone and light-weight subtask, which can make the most of the computational resources and capture much longer-term dependency (i.e. using longer input sequences), resulting in more consistent and natural-looking videos.

3.2. Implicit Representation Parameterization

Unlike previous methods that reconstruct 3D face shapes with the extracted facial parameters [30, 40] and convert them to videos, we map such facial parameters to an implicit representation \( I_F \) and use \( I_F \) to condition the video synthesis. In this way, our framework takes the best of both
explicit and implicit face representation approaches as i) it makes good use of the interpretability of the facial parameter space that facilitates controllability of the synthesis process; ii) it captures more realistic facial features with the high expressive power of the implicit representation; iii) it avoids the unnecessary introduction of facial priors that are inconsistent with ground truth when performing 3D face reconstruction from sparse facial parameters.

As Fig. 2 shows, we implement the mapping between facial parameters and its implicit representation using an EG3D [3] generator. Specifically, our facial parameters consist of identity, expression and pose. For identity and expression, we concatenate them and employ a simple mapping network to map them to an intermediate latent vector \( z \). For pose, we represent it with the camera pose \( R, t, K \) and intrinsic matrix \( K \), and use it to query the 3D positions using the tri-plane structure. Following [3], we feed \( z \) to the EG3D generator as both input and condition vector, and \( R, t, K \) to it as the camera pose, and obtain \( I_F \) as an implicit representation of the input facial parameters. Note that we use face reconstruction as a proxy task (\( I_{\text{face}} \) denotes the reconstructed face) for the training and discard the decoder in the testing stage.

Remark. We use EG3D [3] rather than NeRF [12, 35] as EG3D is a computationally efficient and expressive architecture that supports the generation of high-resolution images in real time and greatly preserves 3D structure.

3.3. Rendering with PIR

As mentioned above, although the implicit representation \( I_F \) carries realistic facial features, its sparse input (i.e., facial parameters) cannot capture the fine details of the input video. To this end, we formulate facial reenactment as a video inpainting problem conditioned on the implicit representation \( I_F \). Specifically, as shown in Fig. 2, given a masked video frame \( I_M \), we first use an encoder \( E \) to extract its feature maps with the same resolution as \( I_F \). Then, we concatenate them with \( I_F \) and feed the concatenated feature maps to the decoder \( D \) to generate the reenactment image \( I \). Skip connections are added between corresponding intermediate layers of \( E \) and \( D \).

Jitter Reduction. Although effective, the proposed rendering method is trained with paired audio and video data, which is not the case during testing. In our experiment (Fig. 3), we observed that new audios may cause slight offsets and deformations of \( I_{\text{face}} \), leading to jitters in the resulting videos. To reduce such jitters, we propose a simple yet effective data augmentation strategy. Specifically, we perturb the camera intrinsic matrix \( K \) with random variables \( x_1, x_2, x_3 \sim U(-s, s) \) when training the rendering
among the components of our framework, audio to expression is a self-contained subtask and can be trained independently, which makes the most of the given computational resources and produces more consistent and natural-looking videos; rendering relies on the result of implicit representation parameterization and is trained afterwards. All of them are trained with short video clips of target persons, including paired audio track and video sequences.

Audio to Expression. Let \( a_i \) be the output feature where \( i = 1, 2, ..., k \) be the \( i \)-th frame of the video, and \( z_{\text{exp, } i} \) be the expression parameter extracted by 3DMM, we train our audio encoder by minimizing the Mean Squared Error (MSE) between them as:

\[
L_{\text{audio}} = \sum_{i=1}^{k} ||a_i - z_{\text{exp, } i}||^2. \tag{2}
\]

\section*{3.4. Training and Loss Functions}

Among the three components of our framework, audio to expression is a self-contained subtask and can be trained independently, which makes the most of the given computational resources and produces more consistent and natural-looking videos; rendering relies on the result of implicit representation parameterization and is trained afterwards. All of them are trained with short video clips of target persons, including paired audio track and video sequences.

Audio to Expression. Let \( a_i \) be the output feature where \( i = 1, 2, ..., k \) be the \( i \)-th frame of the video, and \( z_{\text{exp, } i} \) be the expression parameter extracted by 3DMM, we train our audio encoder by minimizing the Mean Squared Error (MSE) between them as:

\[
L_{\text{audio}} = \sum_{i=1}^{k} ||a_i - z_{\text{exp, } i}||^2. \tag{2}
\]

Implicit Representation Parameterization. We train the mapping between input facial parameters and the implicit representation of reenacted face with a weighted sum of a photometric loss and a perceptual loss [13]

\[
L_{\text{face}} = w_1 ||M_h \odot (I_{\text{face}} - I_{\text{GT}})||^2 + w_2 \sum_i ||\phi_i(I_{\text{face}}) - \phi_i(M_h \odot I_{\text{GT}})||^2, \tag{3}
\]

where \( M_h \) is the head mask, \( \phi_i(*) \) denotes the activation of the \( i \)-th layer in VGG16 [26], \( \odot \) denotes element-wise product operator, \( w_1 \) and \( w_2 \) are the weighting factors.

\section*{Rendering with PIR. To maximize the quality of generated image, we train our rendering network with:

\[
L_{\text{render}} = L_{\text{render}}^{\text{rec}} + L_{\text{render}}^{\text{FM}} + L_{\text{render}}^{\text{GAN}} \tag{4}
\]

where \( L_{\text{render}}^{\text{rec}} \) denotes a reconstruction loss consisting of a weighted sum of a photometric loss and a perceptual loss:

\[
L_{\text{render}}^{\text{rec}} = w_3 ||I - I_{\text{GT}}||^2 + w_4 \sum_i ||\phi_i(I) - \phi_i(I_{\text{GT}})||^2. \tag{5}
\]

Let \( \{D_k|k = 1, 2, 3\} \) be a multi-scale discriminator [32], \( L_{\text{render}}^{\text{FM}} \) denotes a feature matching loss and \( L_{\text{render}}^{\text{GAN}} \) denotes a GAN adversarial loss:

\[
L_{\text{render}}^{\text{FM}} = \sum_{i=1}^{T} \frac{1}{N_i} ||D_k^{(i)}(I) - D_k^{(i)}(I_{\text{GT}})||_1, \tag{6}
\]

\[
L_{\text{render}}^{\text{GAN}} = \sum_k \log D_k(I_{\text{GT}}) + \log(1 - D_k(I)),
\]

where \( T \) is the total number of layers and \( N_i \) denotes the number of elements in each layer. Note that \( L_{\text{render}}^{\text{GAN}} \) is optimized in a minimax manner as those in GAN training.

\section*{4. Experiments}

\subsection*{4.1. Experimental Settings}

\textbf{Dataset.} To achieve high-quality audio-driven facial reenactment, we follow [12, 17] and conduct experiments on three datasets, the HDTF [41], Testset 1 [12], Testset 2 [40]. For HDTF, we selected 8 videos (corresponding to 8 sub-datasets) from it. For Testset 1 and 2, we use the sole video released by the authors for each of them. The time span of these videos is 3-6 minutes. For the test set consisting of unpaired and gender-balanced audio clips, we select 3 from HDTF and collect another 2 Obama audios online. Please note that many previous datasets (e.g., LRW [7], Voxceleb1 [20] and Voxceleb2 [6]) are not suitable as they either have low image quality or consist of many short (a few seconds) video clips of different speakers (e.g., GRID [9] and MEAD [31]), which hinders the generation of high-quality videos and the capture of long-term audio context.
We implement our framework in PyTorch [22] with an Adam optimizer [14]. We train our contextual audio to expression component for 120 epochs with the expression parameters extracted by 3DMM, a context length of $k = 100$ (each frame lasts for 0.04s), and pretrained wav2vec 2.0 weights. Thanks to its well-defined output, we train our audio to expression component simultaneously with the other two components. Our implicit representation parameterization component is trained for 50 epochs using the expression parameters extracted by 3DMM as input and the rendered face $I_f$ as output. The resolution of the resulting parametric implicit representation $I_F$ is $32 \times 64 \times 64$. The three tri-plane features have the resolution of $32 \times 256 \times 256$ and that of output image $I_f$ is $3 \times 512 \times 512$. Our rendering with parametric implicit representation component is trained for another 50 epochs with $s = 3$ for jitter reduction and the augmented $I_F$ whose resolution is $32 \times 32 \times 32$. We use $w_1 = w_2 = w_3 = w_4 = 1$ for Eqs. (3) and (5).

### 4.3. Comparison with the State-of-the-arts

#### 4.3.1 Quantitative Evaluation

We quantitatively compare our method with SOTAs using the following metrics:

- **Lip Movement Accuracy**: We use the Landmark Distance (LMD) [4] to evaluate lip movement accuracy.
- **Lip-sync**: We measure lip synchronization errors with the Audio-Visual Confidence (AVConf) score calculated by SyncNet [8].
- **Sharpness**: We measure frame sharpness with the perceptual-based no-reference objective image sharpness metric (CPBD) [21].
- **Image Quality**: We assess the quality of synthesized video frames by comparing them with the ground truth using Peak Signal to Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM) [33].

As Tab. 1 shows, our method achieves the best or second-best performance for most of the evaluation metrics. For the exceptions, Wav2Lip achieves a higher CPBD score on HDTF but sacrifices the visual quality (blurry mouths with obvious artifacts in Fig. 5) as it only edits the mouth region of the reference images with the remaining part unchanged. In addition, it is trained using a pretrained lip-sync discriminator similar to SyncNet, which “tricks” SyncNet to produce the highest AVConf scores on all three datasets. PC-AVS gets slightly higher AVConf scores than our method on HDTF and Testset 1, but are much worse than ours on all the other metrics, especially LMD. This indicates that

<table>
<thead>
<tr>
<th>Method</th>
<th>HDTF</th>
<th>Testset 1</th>
<th>Testset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM↑</td>
<td>PSNR↑</td>
<td>CPBD↑</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>1</td>
<td>N/A</td>
<td>0.344</td>
</tr>
<tr>
<td>Wav2Lip [24]</td>
<td>0.729</td>
<td>20.352</td>
<td><strong>0.317</strong></td>
</tr>
<tr>
<td>MakeItTalk [44]</td>
<td>0.698</td>
<td>19.956</td>
<td>0.075</td>
</tr>
<tr>
<td>AD-NeRF [12]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FACIAL [40]</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>0.970</strong></td>
<td><strong>36.711</strong></td>
<td><strong>0.305</strong></td>
</tr>
</tbody>
</table>

Table 1: Quantitative comparisons with existing state-of-the-art methods. Since AD-NeRF [12] and FACIAL [40] do not provide pretrained models on the HDTF dataset, we only compare with them on Testset 1 and 2. **Bold**: best results; Underline: second-best results.
PC-AVS learns natural-looking lip movements but fails to capture individual speaking styles. In contrast, our method is more personalized as it takes into account the identity parameters $z_{id}$ and thus excels in the more fine-grained LDM.

### 4.3.2 Qualitative Comparison

We compare our methods with state-of-the-art 2D-based methods, including the explicit ATVG [5] and MakeitTalk [44], and implicit Wav2Lip [24] and PC-AVS [43], in Fig. 5. For 3D-based methods, we compare ours with the explicit FACIAL [40] and implicit AD-NeRF [12] in Fig. 6.

As Fig. 5 shows, our method produces the highest quality results with the most synchronized lip-movement. Specifically, i) ATVG and MakeitTalk fail to produce accurate lip movements as they rely on less expressive 2D facial landmarks; ii) Wav2Lip produces blurry mouths that do not match the sharp parts in the rest of the video frames, making the results unnatural; iii) although PC-AVS produces head movements that are consistent with the ground truth, it cannot well preserve the identity of the speaker. In addition, none of these methods can synthesize high-resolution videos. In contrast, our method allows for the synthesis of high-resolution, high-quality videos with highly synchronized lip movements that preserve facial details well (e.g., teeth and wrinkles), which are crucial for identity preservation and the naturalness of facial reenactment.

As Fig. 6 shows, our method is also superior to AD-NeRF and FACIAL. Specifically, i) AD-NeRF suffers from the artifacts at the head-neck junction which stem from a mismatch between the two NeRFs it uses to model the head and torso separately; ii) FACIAL produces less accurate lip movements due to the less expressive 3D face shape it uses as the intermediate face representation. Please refer to Tab. 2 for a quantitative comparison w.r.t lip movement videos.

### 4.4 Quantitative Evaluation

#### Table 2. Quantitative comparisons of our method with AD-NeRF [12] and FACIAL [40].

<table>
<thead>
<tr>
<th>Method</th>
<th>AVConf↑</th>
<th>Method</th>
<th>AVConf↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD-NeRF [12]</td>
<td>3.607</td>
<td>FACIAL [40]</td>
<td>4.623</td>
</tr>
<tr>
<td>Ours</td>
<td>6.758</td>
<td>Ours</td>
<td>6.678</td>
</tr>
</tbody>
</table>
Figure 6. Comparison with AD-NeRF [12] and FACIAL [40].

Figure 7. Ablation study of our rendering with PIR component. Row 2: the images generated without this component.

Figure 8. Ablation study of our jitter reduction technique.

accuracy between AD-NeRF, FACIAL and our method.

Remark. For the comparison with AD-NeRF and FACIAL (Fig. 6), we use the codes and pretrained models released by the authors and compare the generalizability by feeding all models with 5 unpaired and gender-balanced audio clips mentioned above. The reference images are the corresponding frames of the input audio clips in the new videos.

4.4. Ablation Study

Rendering with PIR. As Fig. 7 shows, without our rendering with PIR component (i.e., use the implicit representation parameterization component to generate the output image $I$ directly, not just the head), the torso changes rapidly within seconds and produces unnatural results. This justifies the necessity of our Rendering with PIR component.

Jitter Reduction. As Fig. 8 shows, without jitter reduction, the rendering component cannot align $I_F$ with $I_M$, thus producing unnatural videos. This justifies the necessity of our jitter reduction technique.

4.5. User Study

We invite 15 volunteers to participate in our user study to evaluate facial reenactment results based on three criteria: lip-sync (audio-lip synchronization), image quality and video realness. We create 3 videos for each method with the same audio input and ask the volunteers to give their ratings on a scale of 1 (worst) to 5 (best) for each video. As Tab. 3 shows, our method scores the highest in all three criteria.

5. Conclusion

In this work, we propose an innovative facial reenactment framework based on our novel parametric implicit representation (PIR). Specifically, our PIR breaks the trade-off between interpretability and expressive power that plagued previous explicit and implicit methods, thus paving the way for controllable and high-quality audio-driven facial reenactment. We have also devised several novel techniques to improve the three components of our framework. Extensive experiments demonstrate the effectiveness of our method.

Acknowledgments

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation (NO. 2020B1515020048), in part by the National Natural Science Foundation of China (NO. 61976250), in part by the Shenzhen Science and Technology Program (NO. JCYJ20220530141211024) and in part by the Fundamental Research Funds for the Central Universities under Grant 22lgqb25. This work was also sponsored by Tencent AI Lab Open Research Fund (NO. Tencent AI Lab RBFR2022009).
References


[40] 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. 1, 3, 5, 6, 7, 8