Photorealistic Audio-driven Video Portraits

Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu

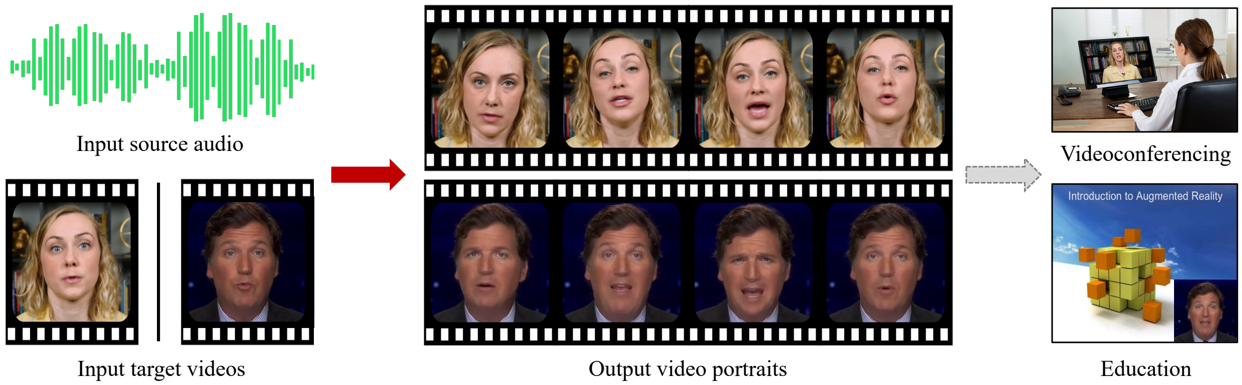


Fig. 1. We present a novel method for generating photorealistic video portraits that correspond to the actor in a target video, reenacted by arbitrary speech audio. Our method has applications in videoconferencing, virtual education and training scenarios.

Abstract—Video portraits are common in a variety of applications, such as videoconferencing, news broadcasting, and virtual education and training. We present a novel method to synthesize photorealistic video portraits for an input portrait video, automatically driven by a person's voice. The main challenge in this task is the hallucination of plausible, photorealistic facial expressions from input speech audio. To address this challenge, we employ a parametric 3D face model represented by geometry, facial expression, illumination, etc., and learn a mapping from audio features to model parameters. The input source audio is first represented as a high-dimensional feature, which is used to predict facial expression parameters of the 3D face model. We then replace the expression parameters computed from the original target video with the predicted one, and rerender the reenacted face. Finally, we generate a photorealistic video portrait from the reenacted synthetic face sequence via a neural face renderer. One appealing feature of our approach is the generalization capability for various input speech audio, including synthetic speech audio from text-to-speech software. Extensive experimental results show that our approach outperforms previous general-purpose audio-driven video portrait methods. This includes a user study demonstrating that our results are rated as more realistic than previous methods.

Index Terms—Audio-driven animation, facial reenactment, generative models, talking-head video generation

1 Introduction

Visual information from a speaker's face, such as their lip movements, can improve speech comprehension in general human communication. It plays a critical role in comprehending speech content for the hearing impaired or when the acoustic signal is corrupted by background noise. In many scenarios, such as telephony or VR/AR professional training for doctors and pilots, however, speech communication is purely acoustical and the visual counterpart is missing due to the lack of cameras, privacy concerns or the limited bandwidth of networks. To improve speech comprehension in these scenarios, many approaches have been proposed to synthesize a talking face from the acoustic speech in real time, as a virtual [21, 28, 47] or a photorealistic avatar [40, 46]. Users' sense of presence is also increased when the avatar is similar to the real user [26, 36].

Video portraits provide photorealistic visual content of a person's face, perfectly maintain their identity, and are commonly used in videoconferencing, virtual anchoring and virtual training. However, it is challenging to generate plausible visual content that matches the acoustic signal, and any misalignment between the mouth motion and the pronunciation can degrade the visual experience. The essential technical issue behind this challenge is the mapping from a raw audio signal to photorealistic imagery. Existing audio-driven video portrait generation techniques produce results that are not sufficiently photorealistic for application requirements, or do not generalize well to given audio or target video inputs. Video portraits can be generated and edited by editing text [3, 18]. However, current text-based methods either focus on cut and transition operations of prerecorded videos [3], which cannot generate results with new text, or synthesize new audio-visual speech content [18] for a specified performer with at least one hour footage for feature searching, and is not suitable for use in real-time applications.

In this work, we present a novel real-time photorealistic video portrait generation method from speech audio. Instead of directly learning to predict the 2D portrait image sequence from audio, we propose to predict the facial expression component of a parametric 3D face model from audio input using neural networks. We then blend the predicted facial expressions with the other components computed from the target video, to generate a reenacted 3D face sequence. Using a neural face renderer, trained on the target video, the reenacted 3D face is converted

- *Xin Wen is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100091, China. E-mail: xinwen@buaa.edu.cn.*
- *Miao Wang is with the State Key Laboratory of Virtual Reality Technology and Systems, Research Institute for Frontier Science, Beihang University, Beijing 100091, China, and also with Peng Cheng Laboratory, Shenzhen 518040, China. E-mail: miaow@buaa.edu.cn.*
- *Christian Richardt is with the with the Department of Computer Science at the University of Bath, UK. E-mail: christian@richardt.name.*
- *Ze-Yin Chen is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100091, China. E-mail: chenzeyin980607@gmail.com.*
- *Shi-Min Hu is with the BNRist, Tsinghua University, Beijing 100084, China. E-mail: shimin@tsinghua.edu.cn.*
- *Corresponding author: Miao Wang.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

to a photorealistic video portrait. All source code is publicly available.[1]

We make the following contributions in this work:

- Given input speech audio, our method generates a photorealistic video portrait of a target actor. A three-minute video of the target is sufficient for training our complete pipeline – much less data than required by existing methods.
- We present an audio to facial expression mapping module that can transform identity-independent speech audio into a target actor's facial expression parameters, only trained on a single target portrait video.
- We evaluate the efficacy of our method with an extensive user study. Our results were rated the most photorealistic by participants when compared to existing general-purpose audio-driven video portrait methods.

2 Related Work

2.1 Monocular 3D Face Reconstruction

Monocular 3D face reconstruction aims to reconstruct facial geometry and appearance, including facial expressions, from visual data [4, 19, 46, 52]. This is the basis for facial reenactment. Model-based methods are the common practice; they employ a parametric face model [4, 17] as a prior when minimizing the reconstruction energy in an *analysis-by-synthesis* paradigm. Based on the type of visual input data, methods can be categorized as *single-image-based* [4, 42], *photo-collection-based* [16, 43] and *video-based* [19, 44, 52]. Recently, various deep-learning-based approaches have been proposed to estimate 3D model parameters [16, 20, 23, 31, 49]. Apart from model parameters, some approaches also regress fine-scale skin details [6, 42, 49]. This is an active area of research with a large variety of works; for more information, we refer to recent surveys [17, 55, 65].

2.2 Video-driven Facial Reenactment

Video-driven facial reenactment takes two face sequences as input: a source and a target. The target face is reenacted using the expression parameters of the source face. *Face2Face* [52] is a real-time video reenactment method that adopts a high-resolution skin texture and synthesizes the mouth cavity using a data-driven approach. Averbuch-Elor et al. [1] proposed a technique to automatically animate a still image portrait using a driving video, by transferring the facial expressions from the video to the image via 2D warping, and synthesizing the mouth interior. Similar portrait image animations can be achieved by few-shot learning from a talking-head video. Zakharov et al. [62] employed meta-learning on a large video dataset as pretraining, and perform few-shot or one-shot learning on unseen people with adversarial training. In *deep video portraits* [30], a generative adversarial network (GAN) [22] is proposed to produce photorealistic video with full control of head pose, facial expression and eye gaze of the portrait. Kim et al. [29] proposed a visual dubbing method which can maintain the signature style of the target actor during talking. Instead of directly replacing the target expression with that of the source actor, this approach learns the mapping in an unsupervised manner with cycle consistency [60, 64].

2.3 Audio-driven Facial Reenactment

The goal of audio-driven facial reenactment is to generate photorealistic video portraits that are in sync with the input audio stream. Chung et al. [12] developed a technique that animates a still image portrait following an audio speech. With both image and audio jointly encoded into a latent space through an encoder network, a decoder network synthesizes the talking head. Both the encoder and decoder are trained in an unsupervised manner. Zhou et al. [63] proposed a method to learn a disentangled audio-visual representation in a novel adversarial training process. This method can take either audio or video to drive the target actor. Chen et al. [11] first transformed audio features to facial landmarks as an intermediate feature, and then generated speech frames conditioned on the landmarks with an attention mechanism. Prajwal et al. [40] proposed a face-to-face translation method that generates talking faces of any person given a speech segment. The LipGAN architecture comprises a generator to synthesize portrait video frames from source audio and target frames, and a discriminator to determine if the synthesized face image is synced with the audio. However, blur and jitter can be observed in their results, because temporal stability of the synthetic content is not guaranteed. While the above four methods can take arbitrary audio as input to reenact arbitrary actors using a single input image, the results are not sufficiently photorealistic due to the low image quality. Vougioukas et al. [54] proposed an end-to-end method to generate talking head videos using a still image and speech audio. A Temporal GAN with three discriminators is employed to achieve sharp frames, audio-visual synchronization, and realistic expressions. *VOCA* [15] is a technique for realistic 3D facial animation from arbitrary audio, based on a new 4D face dataset of twelve speakers.

Suwajanakorn et al. [46] synthesized high-fidelity talking-head videos of former US president Barack Obama, using an audio stream of him. A recurrent neural network (RNN) is trained on 14 hours of his speech to predict the mouth shape from the mel-frequency cepstral coefficients (MFCC) audio feature. A photorealistic mouth region is synthesized within a manually drawn mask using the median texture of retrieved candidate frames. The mouth region sequence is finally composited on the time-warped target video background. Although this approach can synthesize accurate lip-synced video, it requires 14 hours speech video of a specific target identity to train the network, and does not generalize to other identities. Yu et al. [61] proposed a method for generating talking-head video from text and/or audio input. Optical flow and self-attention are introduced to model temporal and spatial dependencies, respectively. However, like Suwajanakorn et al. [46], their method is only demonstrated on US presidents Donald Trump and Barack Obama, and does not generalize beyond them.

We have noticed some concurrent work relevant to ours. Similar to our work, *Neural Voice Puppetry* [50] presents an audio-to-expression network that is trained on a large corpus of TV broadcasts. The lower face is rerendered using the predicted expression from audio with deferred neural rendering [51]. To fill the gap between jaw and neck, an additional standalone inpainting network is employed. In contrast, we address this issue using a simple mask expansion process that is controlled by a facial expression parameter and thus more efficient computationally. Song et al. [45] proposed an ID-removing network to predict expression parameters, and a universal translation network that transforms landmark heatmaps to photorealistic video for arbitrary targets. However, using landmark heatmaps as input to the neural face renderer can introduce jitter, as it is challenging to maintain the temporal coherency of landmarks. More recently, Yi et al. [59] proposed a personalized learning-based head pose generation method to enhance the fidelity of talking-head videos. Less data (about 10 seconds) is required to train an image translation network through a memory-augmented GAN. However, due to errors in their face reconstruction, the reconstructed face sequence is unstable, which harms GAN convergence. Noticeable artifacts are also visible in the mouth cavity, which reduces the fidelity and user experience.

2.4 Deep Generative Models and Neural Rendering

Recently, GANs have been proposed for image synthesis from noise. This approach can be extended with a conditional input setup [35], which is usually used to bridge the gap between two different but relevant domains. The *pix2pix* image-to-image translation method [27] is widely regarded as one benchmark method of conditional GAN-based image synthesis. This paradigm can be extended to video-to-video translation to synthesize video frames with temporal coherency. Wang et al. [57] proposed a method to generate high-resolution and temporally smooth video in a course-to-fine manner with a recurrent network. The *Recycle-GAN* approach [2] enables unpaired learning of a coherent video-to-video translation. Few-shot video-to-video translation [56] learns to synthesize videos of unseen subjects via a novel network weight generation module. Video-to-video translation shows impressive results in many applications, especially for face reenactment, visual dubbing [18, 29, 30] and even full-body reenactment [10, 33].

Nowadays, many approaches combine the power of neural net-

[1] https://github.com/xinwen-cs/AudioDVP

works and traditional rendering using neural rendering [48]. Neural textures [51] are a novel learnable component, which mimic texture maps used in the traditional graphics pipeline. They show compelling results in applications of novel view synthesis, scene editing and animation synthesis. Meshry et al. [34] trained a neural rerendering network which takes a deep framebuffer consisting of depth, color and semantic labeling as input and outputs realistic renderings of the scene under multiple appearances. Thies et al. proposed a learning-based image-guided rendering technique [53] that combines image-based rendering and GAN-based image synthesis. This method can generate photorealistic rerenderings of reconstructed objects for virtual and augmented reality applications, such as virtual tours, showrooms and sightseeing. In our method, a neural face renderer is employed to translate the rough rendering of the lower face to photorealistic imagery.

3 Audio-driven Video Portrait Generation

Given a source speech audio, our method aims to generate a photorealistic video portrait for a given target video. To achieve this goal, we employ a 3D face rig to bridge the gap between the raw input audio and photorealistic output video modalities. This intermediate model avoids overfitting to spurious correlations between the audio and visual signals. The pipeline of our method consists of three main components, as illustrated in Figure 2: monocular 3D face reconstruction, audio-to-facial-expression mapping ('*Audio2Expression*'), and neural face rendering. Given a target video $\mathcal{V}_t$, we first reconstruct a parametric 3D face model with expression, geometry, texture, pose and illumination parameters for every frame (Section 3.1). From the same video, we learn a mapping from audio features to facial expression parameters of the same parametric 3D face model (Section 3.2); this mapping can transform a speech audio – even from other people – to the expression parameters of the target actor. For a source audio track $\mathcal{A}_s$, our method predicts expression parameters from the audio, blends the predicted expression parameters with the face model reconstructed from the target video, and rerenders the audio-driven face images of the target actor. As can be seen, these rerendered images are not photorealistic. To tackle this issue, we train a neural face renderer to translate the rendered lower face regions to photorealistic ones that are composited into the original target video frame as the final result (Section 3.3).

3.1 Monocular 3D Face Reconstruction

For the target video $\mathcal{V}_t = \{I_1, \ldots, I_M\}$ with M frames, we first track the face in all frames and register a 3D face model. Let $\{\mathcal{X}_1, \ldots, \mathcal{X}_M\}$ denote the sequence of face model parameters that fully describe the facial performance of the target video $\mathcal{V}_t$. We follow the single-image-based method by Deng et al. [16] and adapt it to video-based 3D face reconstruction. In this section, we first briefly introduce the parametric face model we use. Then, we describe the image formation process to transform the 3D face model into a 2D image. Finally, we discuss the energy terms used for face model fitting.

3.1.1 Parametric Face Model

We use a 3D morphable model (3DMM) to represent the face [4, 17]. The 3DMM consists of a template triangle mesh with N_v vertices and an affine model that defines the facial geometry $v \in \mathbb{R}^{3N_v}$ (stacked 3D positions of vertices) and the stacked per-vertex diffuse reflectance $r \in \mathbb{R}^{3N_v}$ in terms of the coefficients $\{\alpha_k\}$ for geometry, $\{\delta_k\}$ for expressions, and $\{\beta_k\}$ for reflectance (color):

$$\begin{aligned} v(\alpha,\delta) &= a_{\text{geo}} + \sum_{k=1}^{N_\alpha} \alpha_k b_k^{\text{geo}} + \sum_{k=1}^{N_\delta} \delta_k b_k^{\text{exp}}, \\ r(\beta) &= a_{\text{ref}} + \sum_{k=1}^{N_\beta} \beta_k b_k^{\text{ref}}. \end{aligned} \tag{1}$$

Here, the vectors $a_{\text{geo}}, a_{\text{ref}} \in \mathbb{R}^{3N_v}$ represent the average facial geometry and reflectance, respectively, $\{b_k^{\text{geo}}\}_{k=1}^{N_\alpha}$ is the geometry basis, $\{b_k^{\text{exp}}\}_{k=1}^{N_\delta}$ is the expression basis, and $\{b_k^{\text{ref}}\}_{k=1}^{N_\beta}$ is the reflectance basis, all computed from facial scan data using principal component analysis (PCA). We adopt the 2009 Basel face model [39] for the facial geometry (a_{geo}, b^{geo}) and reflectance (a_{ref}, b^{ref}), and augment it with the facial expressions b^{exp} from Guo et al.'s coarse-to-fine learning framework [23], which builds on FaceWarehouse [7]. We use $N_\alpha = 80$, $N_\delta = 64$ and $N_\beta = 80$. The rigid head pose is represented by rotation $R \in \mathrm{SO}(3)$ and translation $T \in \mathbb{R}^3$.

3.1.2 Image Formation Process

To render the 3D face model $\mathcal{X}$ as a synthetic image $\hat{I}$, we furthermore need to model the illumination and the camera. We assume a Lambertian surface and distant scene illumination to approximate environment lighting using spherical harmonics (SH) [41]: $C(r_i, n_i, \gamma) = r_i \odot \sum_{b=1}^{B^2} \gamma_b Y_b(n_i)$, where B is the number of SH bands, $\gamma_b \in \mathbb{R}^3$ are the RGB SH coefficients, $Y_b \colon \mathbb{R}^3 \to \mathbb{R}$ are SH basis functions, r_i and n_i are the reflectance and unit normal vectors of vertex i, respectively, and '$\odot$' is the element-wise product. We choose $B = 3$ bands of SH, with $B^2 = 9$ coefficient vectors, resulting in the SH illumination coefficients $\gamma \in \mathbb{R}^{27}$. Our complete face model can be represented by a vector $\mathcal{X} = (\alpha, \delta, \beta, \gamma, R, T) \in \mathbb{R}^{257}$.

We model the virtual camera as a pinhole camera with a perspective projection $\Pi \colon \mathbb{R}^3 \to \mathbb{R}^2$, which maps 3D points from camera space to 2D image space. For a vertex $v_i \in v(\alpha, \delta)$ of a model $\mathcal{X}$, we compute its image-space coordinates $u_i(\mathcal{X})$ and corresponding color $c_i(\mathcal{X})$ using the aforementioned illumination and camera model. Finally, $\{u_i(\mathcal{X})\}_{i=1}^{N_v}$ and $\{c_i(\mathcal{X})\}_{i=1}^{N_v}$ are fed into a differentiable rasterizer to generate the rendered synthetic image $\hat{I}(\mathcal{X}, \Pi)$. In addition to Genova et al. [20], our rasterizer is implemented with CUDA to gain GPU acceleration, which can speed up both training and inference.

3.1.3 Model Fitting

We use a ResNet-50 network [25] pretrained on VGGFace2 [9] to estimate the face model parameters $\mathcal{X}$ from an input image I, as we found it to produce temporally more coherent results than direct optimization. Specifically, we modify the final fully-connected layer of the network to have 97 dimensions (without geometry and reflectance; see below), and adopt an analysis-by-synthesis approach that minimizes the discrepancy between a synthetic rendering of the model and the input image. The reconstruction loss combines three terms: dense photometric alignment, sparse landmark alignment, and statistical regularization.

We measure the photometric discrepancy between the input frame I and the synthetic image $\hat{I}$ rendered from the model $\mathcal{X}$ using a photo-consistency loss computed across all pixels i in the face region $\mathcal{M}$:

$$\mathcal{L}_{\text{photo}}(\mathcal{X}) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \left\| I(i) - \hat{I}(i) \right\|_2 . \tag{2}$$

We use a sparse landmark alignment constraint to encourage landmarks on the 3D mesh to project close to the corresponding detected 2D landmarks in the input image. We detect $N_L = 68$ landmarks $\{s_1, \ldots, s_{N_L}\}$ in each video frame using an off-the-shelf face alignment network [5], and compute the sparse landmark alignment loss as the weighted Euclidean distance between projected landmarks $u_{\tau_i}(\mathcal{X})$ and detected landmarks s_i:

$$\mathcal{L}_{\text{land}}(\mathcal{X}) = \frac{1}{N_L} \sum_{i=1}^{N_L} \omega_i \left\| u_{\tau_i}(\mathcal{X}) - s_i \right\|_2 . \tag{3}$$

Here, τ_i is the vertex index of the 3D face model corresponding to landmark i in image space, and ω_i is a landmark-specific weight set to 50 for the 20 mouth and 12 eye landmarks, and otherwise set to 1.

To prevent the degeneration of face shape and reflectance, we further employ a regularization loss $\mathcal{L}_{\text{reg}}(\mathcal{X})$ on the regressed 3DMM coefficients, which enforces a prior towards the mean face under Gaussian distribution [16, 49].

The total model-fitting loss is defined as:

$$\mathcal{L}(\mathcal{X}) = \lambda_{\text{photo}} \mathcal{L}_{\text{photo}}(\mathcal{X}) + \lambda_{\text{land}} \mathcal{L}_{\text{land}}(\mathcal{X}) + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}(\mathcal{X}), \tag{4}$$

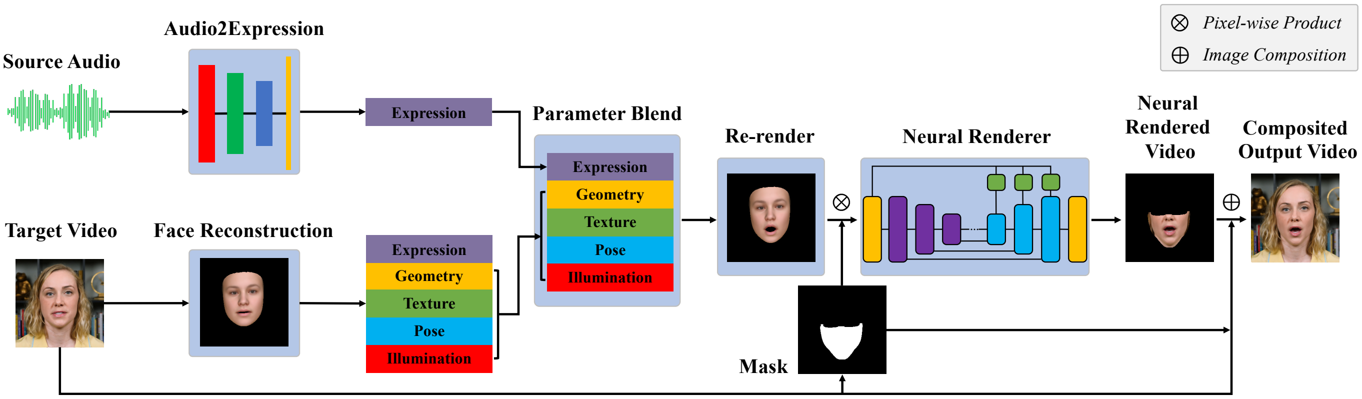


Fig. 2. Pipeline of our approach. From left to right: First, we estimate the parameters of a 3D face model for the target video portrait via monocular face reconstruction (Section 3.1), and compute the facial expression parameters from the source speech audio (Section 3.2). Next, we create a new face by blending the facial expression parameters predicted from the source audio with the other parameters from the target video, and rerender synthetic images of the new face model. Finally, we use a neural face renderer to generate photorealistic renderings from the synthetic images, and composite the result on top of the dynamic video background (Section 3.3).

Table 1. Architecture of our Audio2Expression mapping network.

Type	Kernel	Stride	Outputs
Conv1D	3	1	5×254
Conv1D	3	1	3×252
Conv1D	3	1	1×250
FC	–	–	64

where we use $\lambda_{\text{photo}} = 1.9$, $\lambda_{\text{land}} = 0.0016$, and $\lambda_{\text{reg}} = 0.0003$ for all experiments. For the detailed derivations, we refer to Genova et al. [20]. Before fitting the model to the full target video, we randomly select 8 frames to regress the geometry and reflectance parameters of each actor and keep them constant. We then train our face reconstruction network for 20 epochs on the target video with a batch size of 5 and a learning rate of 2×10^{-5}. While this subsection provides technical details of face fitting with implementation differences compared to previous work, we clarify that this is not one of our main contributions.

3.2 Audio to Facial Expression Mapping

To reenact the face model based only on an audio stream, we next introduce a facial expression mapping method that estimates facial expression parameters of the face model from the input audio. First, we use *AT-net* [11] to robustly extract high-level features from audio. *AT-net* was originally designed for creating landmark animation from an audio stream, and was trained on the LRW dataset [13], a large-scale lip reading corpus based on BBC broadcasts. To obtain high-level audio features, we convert the input audio stream into MFCC features, which we feed into *AT-net* and take the 256-D output feature of the ante-penultimate layer as the robust high-level features. We found that these features are effectively independent of any specific identity and contain sufficient information for expression prediction. As a result, we extract a 256-D feature vector F for every 40 ms segment of the input audio $\mathcal{A}_s$ (corresponding to one video frame at 25 frames per second).

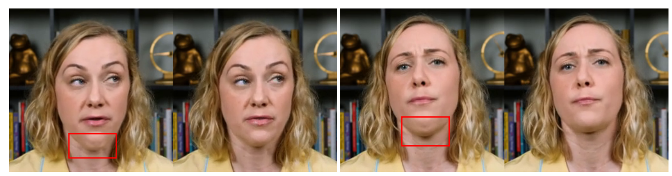
We propose an audio-to-facial-expression mapping network $\mathcal{H}$ that takes these audio features as input and predicts expression parameters. To maintain temporal coherency, for each time step t, we stack audio features as inputs along the timeline within a sliding window, and get $\mathbf{F}_t = \{F_i\}_{i=t-N_w}^{t+N_w}$, where $N_w = 3$ is the radius of the sliding window. We set non-existing prior or subsequent features F to zero. We use three layers of 1D convolutions to integrate space-time information, and a fully-connected layer with 64 nodes to output the predicted expression coefficients. The network structure is given in Table 1.

We use the mean squared error (MSE) loss L_{exp} to train $\mathcal{H}$:

$$\mathcal{L}_{\text{exp}} = MSE(\mathcal{H}(\mathbf{F}_t) - \delta_t), \tag{5}$$

where δ_t is the expression parameter at time step t obtained from the reconstructed target video. We train the network using Adam [32] with default settings for 10 epochs with a batch size of 5.

Our audio to facial expression mapping method is only trained on the target video $\mathcal{V}_t$ (typically three minutes long), and is capable to transform the speech audio from an arbitrary person to the facial expression parameters of the target actor.



(a) (b) (c) (d)

Fig. 3. Effect of mask expansion for neural face rendering. **(a)** and **(c)**: directly compositing the synthetic face region within the original lower face mask into the target video can result in dual jaw artifacts. **(b)** and **(d)**: with our face mask expansion, the synthetic lower face and partial neck are both composited into the target video, which avoids conflicting content around the jaw region.

3.3 Neural Face Renderer

We combine the expression parameters estimated from the source audio $\mathcal{A}_s$ with the geometry, reflectance and illumination reconstructed from the target video, to rerender the face model via the image formation process and obtain a sequence of synthetic face images. However, the synthetic images clearly look computer-generated and not photorealistic. To make the synthetic faces more photorealistic and natural-looking, we employ a neural face renderer to translate synthetic renderings into photorealistic images.

Before applying neural face rendering to the synthetic face rendering, we introduce a masking strategy to distill the lower face region with a predefined mask that covers jaw, mouth and part of the nose. We use the neural face renderer to predict the content only within the mask, and composite the predicted content with the target video to produce the final result. With this masking strategy, the training is focused on the mouth animation of the lower face, and avoids the instability of any dynamic background of the target video. At first, we extract a raw mask as follows: we mark all face vertices with y-coordinates less than the threshold $\xi = 0$ (assuming normalized model space coordinates in $[-1, +1]$). We rasterize the masked face to get the binary lower face mask for every frame. However, directly compositing the predicted content into the target video can introduce a doubled jaw due to the inconsistency between predicted expression parameters and the original ones in the target video, as shown in Figure 3. Inspired by InverseFaceNet [31], we explicitly expand the mask around the jaw region by decreasing the value of the first component of the expression

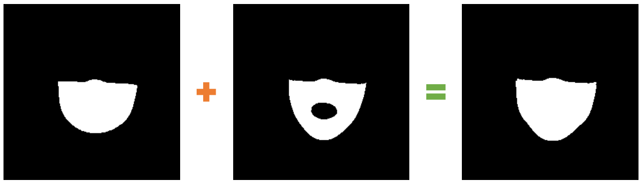


Fig. 4. Expansion for the lower face mask. We generate a modified mask with the mouth fully open (middle) by setting the value of the first expression parameter to "–8", which makes the jaw cover part of the neck. The expanded mask is the union of both masks.

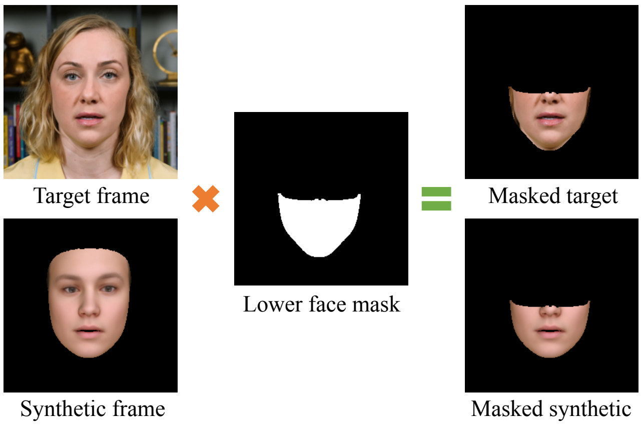


Fig. 5. We build a paired training corpus by applying the expanded mask to each target video frame and synthetic frame, respectively.

parameters to fully open the mouth, so that the jaw covers part of the neck. This process is illustrated in Figure 4.

We create a paired training corpus for the neural face renderer by applying the expanded mask to each target video frame I and the corresponding synthetic image $\hat{I}$, as shown in Figure 5. To avoid excessive notation, let the training corpus $\{(I_t, \hat{I}_t)\}_{t=1}^{M}$ denote the content of images within the expanded lower face mask (instead of complete images). The neural face renderer learns to convert the synthetic rendering to a photorealistic one for the target actor. Following *deep video portraits* [30], we train a neural face renderer consisting of a U-Net-based generator $\mathcal{G}$ and a discriminator $\mathcal{D}$ that are optimized alternatively in an adversarial manner. The generator comprises an encoder and a decoder. The encoder repeatedly downsamples the input tensor using a convolutional layer, followed by batch normalization and a leaky ReLU. The decoder synthesizes high-quality output from the low-dimensional latent representation by upsampling using transposed convolution, batch normalization, dropout and ReLU. The discriminator employs a PatchGAN [27]. For the full network architecture, please refer to *deep video portraits* [30] and our source code. The input of the generator $\mathcal{G}$ is a stacked tensor $\mathbf{T}_t = \{\hat{I}_t\}_{i=t-Nw}^{t+Nw}$, which is composed in the same manner of Section 3.2 to maintain temporal coherency. The input of the discriminator $\mathcal{D}$ is $\mathbf{T}_t$ combined with either the ground-truth image I_t or the neural rendered image $\mathcal{G}(\mathbf{T}_t)$. The optimal network parameters of the generator $\mathcal{G}$ can be obtained by solving following problem:

$$\mathcal{G}^* = \arg\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{G}, \mathcal{D}). \tag{6}$$

The full training objective consists of a photometric reconstruction loss $\mathcal{L}_r$ and an adversarial loss $\mathcal{L}_{\text{adv}}$, weighted by $\lambda = 100$:

$$\mathcal{L}(\mathcal{G}, \mathcal{D}) = \mathcal{L}_{\text{rec}}(\mathcal{G}) + \lambda \mathcal{L}_{\text{adv}}(\mathcal{G}, \mathcal{D}). \tag{7}$$

The photometric reconstruction loss $\mathcal{L}_{\text{rec}}$ encourages the sharpness of the synthesized output and can be formulated as:

$$\mathcal{L}_{\text{rec}}(\mathcal{G}) = \|I_t - \mathcal{G}(\mathbf{T}_t)\|_1. \tag{8}$$

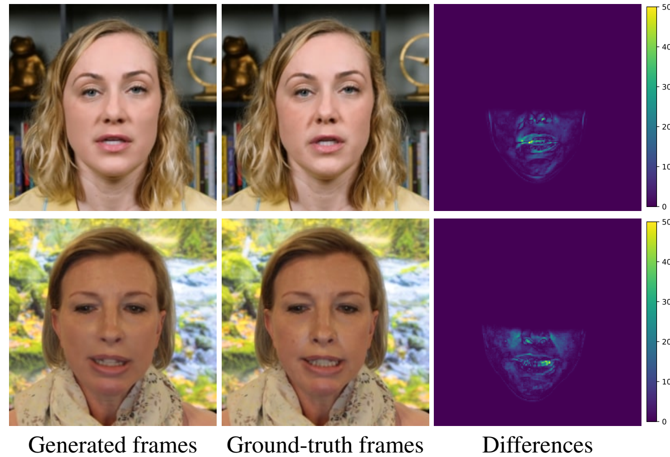


Fig. 6. Visualization of absolute pixel-wise differences between our generated results and ground-truth frames (in range [0, 255]).

The vanilla GAN adversarial loss is:

$$\mathcal{L}_{\text{adv}}(\mathcal{G}, \mathcal{D}) = \log \mathcal{D}(I_t) + \log(1 - \mathcal{D}(\mathcal{G}(\mathbf{T}_t))). \tag{9}$$

We train the network using the Adam optimizer with default settings [32]. We train our networks from scratch with weights initialized following a normal distribution $\mathcal{N}(0, 0.02^2)$. The training process takes 250 epochs with a batch size of 16 and learning rate of 0.0002.

At inference time, we composite the output of our neural face renderer with the background of the target frame using a Gaussian-smoothed lower face mask, as illustrated in Figure 2.

4 Experiments

We demonstrate our audio-driven video portrait generation approach by performing qualitative and quantitative evaluations. We encourage readers to watch our supplementary video for results in action.

Datasets. We test our approach on a set of 11 target videos that were collected from YouTube and prior work [29]. Table 2 provides a summary of these videos, including their lengths and languages. The average length of video clips is 3 minutes. In a preprocess, we align all video frames using the detected landmarks to ensure that the upper body occupies the main space of the image. The aligned frames were further cropped and resized to 256×256 pixels.

Implementation Details. All networks were implemented in PyTorch [38]. We implemented the rasterizer [20] with CUDA acceleration and integrated it into PyTorch. All experiments are conducted on a computer with a 3.6 GHz CPU, 32 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU.

Runtime Performance. For a 3-minute target video portrait, it takes about 30 minutes to reconstruct the face model, 20 seconds to train the Audio2Expression module, and 6.5 hours to train the neural face renderer. In the online testing stage, it takes 2 ms to predict expression parameters from audio, 3 ms to rerender the face model, 13 ms to perform neural face rendering, and 2 ms to composite the neural rendered face region into the target frame. In summary, it takes 20 ms to generate one frame of a video portrait from audio, which is sufficient for real-time applications (50 Hz).

4.1 Video Portrait Results

Self-reenactment. We evaluate our method by using the audio from the target video as input (i.e., *self-reenactment*), and compare the generated video portrait with the ground-truth target video. We perform this test on 60-second test videos of actors **A** and **B** with ground-truth head poses. We calculate the absolute average pixel-wise differences between generated frames and ground-truth frames for each channel in RGB color space in $[0, 255]$. The self-reenactment results with visualizations of errors for two actors are shown in Figure 6. The average difference between generated frames and the ground truth, on

Table 2. List of datasets used in our results and comparisons. The lengths of training segments are provided in seconds (s).

Name	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
Thumbnail											
Language	English	English	German	English	German	English	English	English	English	English	English
Length	180 s	240 s	180 s	240 s	180 s	240 s	87 s	80 s	180 s	240 s	180 s

Fig. 7. Self-reenactment of a reading child (video **K**). Please refer to the supplementary video for more details and video examples.

the whole test set of 1500 frames for videos **A** and **B** is 7.95 and 5.86, respectively. Corresponding videos are in the supplement.

Results for Children. Our method can handle video and audio of children. This is because the Audio2Expression and face model fitting modules in our method are robust to both adults and children. In Figure 7, we show representative frames of self-reenactment for a storytelling video of a child. We also use the child's voice to reenact adult portraits. Corresponding videos are in the supplement.

German Speech Results. Although the high-dimensional feature extraction from audio in our Audio2Expression network is only trained on an English corpus, it generalizes to some other languages, such as German. In Figure 8, we present two video portrait results that use the German audio tracks from videos **C** and **E** to drive and reenact the target actors in videos **B** and **D**, respectively, which were originally speaking in English. Our approach generates plausible results on German speech and can be extended to a multilingual setup. For the full results, we refer to the supplementary video.

Multi-target Video Portraits. Our method can even generalize to reenacting a target video portrait using arbitrary speech audio from different people. We show an example of using one source audio to reenact multiple target videos in Figure 9: President Obama's speech is used as the source audio to reenact four target videos of different actors as well as another Obama video. As can be observed, our method creates plausible mouth expressions for both Obama himself and other actors synchronized to the source audio. Corresponding video results are provided in our supplementary video.

Synthetic Audio. Our method can also take synthetic speech audio as input and generate plausible speech video portraits. To demonstrate the effectiveness of this approach, we use synthesized audio generated by the text-to-speech tool WaveNet [37]. We select the first paragraphs of the novel *A Tale of Two Cities* by Charles Dickens and the poem *Youth* by Samuel Ullman as scripts, and generate speech audios with a female and a male narrator, respectively. The audios are used to drive the speech of target actors in videos **B** and **J**. We include the results in our supplementary video. This expands the range of application scenarios, as audio can be obtained from various off-the-shelf text-to-speech tools.

4.2 Visual Comparisons

We perform comparisons to state-of-the-art audio-driven video portrait generation methods. We first compare our method with the 2D-based methods DAVS [63], ATVG [11] and LipGAN [40] that directly predict video portrait frames without 3D face modeling. The LipGAN method can take both an image or a video of the target actor as input (henceforth denoted "LipGAN (img.)" and "LipGAN (vid.)", respectively). We use part of each video (**A**, **B** and **I**) for training, and the remainder of the same video for testing. Both the source audio and target video are taken from the testing segment, without any temporal overlap. Figure 10

Audio **C** → Video **B** Audio **E** → Video **D**

Fig. 8. German speech results. We use a German audio track to generate video portraits of people originally speaking in English. **Left:** the audio from **C** is used to reenact a video portrait for **B**. **Right:** the audio from **E** is used to reenact a video portrait for **D**. Three representative frames are shown from top to bottom, in each column.

shows the comparison results. In DAVS, ATVG and LipGAN (img.), which only take a target image as input, the head pose of the video portrait is fully static and looks unnatural. The lip motions in DAVS and ATVG do not follow the corresponding source audio precisely; moreover, the image quality is reduced. The mouth shape of LipGAN (vid.) is generally better than other alternative methods; however, the details of the mouth interior can be blurry and unstable over time. In contrast, our method generates more natural-looking results with precise mouth synchronization to the audio, higher quality mouth interiors and better temporal stability. We further compare our method to the recent GAN-based speech-driven animation method SDA [54] that takes audio and a still image as input, and outputs an animated face. We ran the authors' implementation, pretrained on the GRID [14], TCD-TIMIT [24] and CREMA-D [8] datasets. As shown in Figure 11 and our supplementary video, our results are of higher visual quality than SDA's results.

We also compare our method with Audio2Obama [46], an audio-driven method that predicts mouth shape from audio features and uses a reconstructed 3D face model to synthesize mouth textures. We show qualitative results of Audio2Obama trained on 14 hours and on 3 minutes of speeches in Figure 12. Although Audio2Obama composites the synthetic mouth region sequences into a time-warped target video to improve the coherency of facial expression and head pose, the mouth shapes in their results are not always consistent with the audio. Further, Audio2Obama is tailored for only one target – Barack Obama, and generally requires hours of consistent training videos. Our technique, instead, is trained on a single video for each target (about 3 minutes in length) and can work with speech audio from other people. For video results, please see our supplementary video.

4.3 Quantitative Evaluation

We further carried out quantitative evaluation compared to the aforementioned 2D-based methods. We calculated SSIM scores [58] between generated results of competing methods and the ground truth for video **A**, **B** and **I**. We also performed an ablation study of our method with

Fig. 9. Multi-target video portraits. **Left:** two representative frames corresponding to the source audio. **Right:** audio-driven video portrait results for multiple actors. The mouth shapes are synchronized well to the source frames.

Fig. 10. Comparison to DAVS [63], ATVG [11] and LipGAN [40] on target videos **A**, **B** and **I**. **Top:** sampled video frames corresponding to the source audio tracks. **Second row to the bottom:** corresponding video portrait frames from different methods. Our approach generates more natural-looking results with precise mouth synchronization to the audio, higher quality mouth interior and better temporal stability.

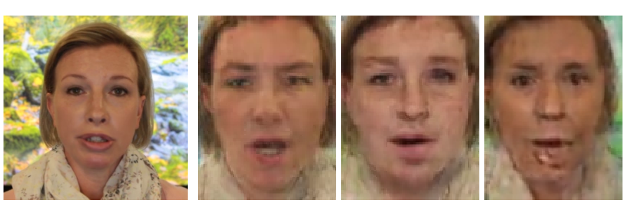


Fig. 11. Comparison to SDA [54]. From **Left** to **Right**: our result and SDA trained on the GRID [14], TCD-TIMIT [24], CREMA-D [8] datasets.

Table 3. Quantitative evaluation of our method and 2D-based methods on the test sets of videos **A**, **B** and **I** using SSIM.

Methods	**A**	**B**	**I**
DAVS [63]	0.5261	0.5806	0.6226
ATVG [11]	0.5720	0.6284	0.6822
LipGAN (img.) [40]	0.5545	0.6135	0.6634
LipGAN (vid.) [40]	0.9451	0.9440	0.9449
Ours w/o neural face rendering	0.9743	0.9732	0.9658
Ours	**0.9858**	**0.9842**	**0.9754**

(a) Comparison to *Audio2Obama* 14 *h* (b) Comparison to *Audio2Obama* 3*m*

Fig. 12. Comparisons to Audio2Obama [46]. Source audio from another Obama speech video is used to drive and reenact the target video portrait. **(a)** Comparison of our model trained on the 87-second video **G** with Audio2Obama trained on 14 hours of speech videos; **(b)** Comparison of our model trained on the 80-second video **H** with Audio2Obama trained on 3 minutes video. **Top:** sampled video frames corresponding to the source audio. **Middle:** corresponding video portrait frames from Audio2Obama. **Bottom:** corresponding video portrait frames from our method. Their results are not always consistent with the driving audio, with the mouth being open or closed when it should not be (see red arrows).

and without the neural face rendering. As shown in Table 3, our method quantitatively outperforms alternative ones.

4.4 User Study

We performed an extensive web-based user study to evaluate our results. We produced short English video clips (6–8 seconds) of size 256×256 from the test dataset in Table 2 using our method and state-of-the-art methods. The videos show a range of expressions, including sarcastic (video **A**), smiling (**B** and **D**) and solemn (**I** and **J**). We conducted subjective rating tasks on each video with 5-point Likert scales.

Participants. We recruited 72 anonymous participants (26 female and 46 male), with an average age of 24.63 years (SD=6.94).

Data and Methods. We generated video portraits with the source audio taken from one segment and the target video (or image) from another non-overlapping segment, both in the same video, by each of the methods, including ours, DAVS [63], ATVG [11], LipGAN (img.) and LipGAN (vid.) [40]. As the output frames in ATVG [11] were severely cropped, we pasted the ATVG results back into input frames for fair evaluation. We also included the original segment corresponding to each audio as test data. This resulted in 30 videos for evaluation (5 videos × 6 methods, including the original segment). We further evaluated Audio2Obama [46] using two videos provided by the authors; one was trained on 14 hours of Obama's speeches, and the other was trained on 3 minutes of Obama speech videos (see Figure 12). Accordingly, we used the same source audio to produce video portraits with our method; however, we only trained on short videos (87 seconds and 80 seconds, respectively). This resulted in 4 Obama videos for evaluation. In summary, 34 video portraits were collected for subjective evaluation.

Procedure. Our web-based user study welcomed participants with a general introduction to the user study on the starting page. Next, participants were asked to fill in information about their age and gender. Before starting the formal study, we showed a test video with the statement "This video clip looks real to me", and corresponding choices from "−2" (strongly disagree) to "+2" (strongly agree), as a warm-up and audio/video check. In the main study, participants played and rated 34 videos one-by-one in random order. Participants could replay each video many times before rating it. After the user study, participants were thanked. The whole process took on average 341 seconds (SD = 120 seconds).

Table 4. User study results in response to the statement "This video clip looks real to me", from "–2" (strongly disagree) to "+2" (strongly agree). Each row lists the percentage of user choices for each rating, the percentage of user choices that agree with the statement (scores "+1" and "+2"), and the mean score for each method. **Top:** Average of 5 video clips (**A**, **B**, **D**, **I** and **J**). **Middle:** Audio2Obama [46] trained on 14 hours of speeches versus our method. **Bottom:** Audio2Obama trained on a 3-minute video versus our method.

Methods	−2	−1	0	+1	+2	agree	mean score
DAVS [63]	68.6	20.0	7.8	2.2	1.4	3.6	−1.52
ATVG [11]	48.9	27.2	12.8	8.9	2.2	11.1	−1.11
LipGAN (img.) [40]	56.6	20.0	15.3	7.5	0.6	8.1	−1.25
LipGAN (vid.) [40]	22.2	31.1	25.3	18.3	3.1	21.4	−0.51
Ours	5.8	16.1	26.7	29.2	22.2	51.4	0.46
Original	1.4	5.0	9.4	29.2	55.0	84.2	1.31
Audio2Obama 14h [46]	1.4	12.5	8.3	40.3	37.5	77.8	1.00
Ours (87 seconds)	6.9	20.8	25.0	30.6	16.7	47.3	0.29
Audio2Obama 3m [46]	7.0	19.4	33.3	34.7	5.6	40.3	0.13
Ours (80 seconds)	9.7	16.7	23.6	36.1	13.9	50.0	0.28

Results. Table 4 summarizes the user ratings in response to the study. The results show that, on average, our method (51.4% ratings agree with "This video clip looks real to me"; average score 0.46) clearly outperforms DAVS [63] (3.6%; −1.52), ATVG [11] (11.1%; −1.18) and LipGAN [40] (8.1%/21.4%; −1.25/−0.51). The average scores of the above competing methods were all negative, while our results were rated on average 0.46 points (within the range [–2, +2]), and more than half of the ratings (51.4%) agreed that our results look real. The original segments corresponding to the source audios gained on average 1.31 points with 84.2% rating them as real. Paired t-tests at the 5% significance level confirm significant differences ($p < 10^{-27}$) between our method and each of the competing alternatives.

The subjective ratings of Audio2Obama [46] and our method reveal that the full Audio2Obama approach trained on 14 hours of videos ("*Audio2Obama 14h*") was considered significantly ($t(71) = 4.81$, $p = 4.1 \times 10^{-6}$) more realistic (77.8% agreement, average score 1.00) than our method trained on just 87 seconds video (47.3% agreement, average score 0.29). However, our method trained on just 80 seconds of video (50.0% agreement, average score 0.28) performed better than Audio2Obama approach trained on 3 minutes of videos ("*Au-*

(a) (b)

Fig. 13. Failure cases. Artifacts can be caused by **(a)** incompatible mouth motion and head pose, or **(b)** exaggerated facial expressions for a target.

dio2Obama 3m"; 40.3% agreement, average score 0.13); the difference was not significant ($t(71) = 1.14$, $p = 0.12$). Nevertheless, our method—once trained on a target actor—can be reenacted by others' audios, and generally requires less training data, which makes it more practical.

5 Discussion

In this work, we have demonstrated photorealistic audio-driven video portrait results for a variety of sequences. While the widely used virtual humans and avatars in VR and AR may have limited lifelikeness, identification preservation and audio-expression synchronization, our audio-driven video portraits are photorealistic, in sync with audio, and maintain the identity of the target actor. Our approach makes a step towards the simplification of photorealistic virtual avatar creation and animation by enabling the reenactment of existing videos with new speech audios. This is especially useful when the network bandwidth is limited or a video capturing device may not be available in VR applications. Nevertheless, our approach has a few limitations that can be addressed in future research.

Our method requires an approximately 3-minute portrait video as training data to generate visually plausible results for each target actor. This is because our pipeline includes a person-specific neural face renderer that needs sufficient training data for each target. An acquisition of a selfie video as short as 30 seconds would be desirable for future daily applications. This introduces an interesting future work that massively reduces the data required to train the renderer, perhaps using meta-learning [56].

In our neural face renderer, only the lower face region is rerendered and integrated into the original face of the target video, which may lead to unnatural artifacts when the original head pose and the source audio are incompatible. For example, Figure 13(a) shows a frame of unnatural moving head pose with a closed mouth. This could be ameliorated with dynamic time warping [46], which retrieves frames from the target video to better align the mouth motion. In addition, full-frame video synthesis, as well as audio-driven head pose prediction are interesting future research directions. Moreover, when the predicted expression parameter is exaggerated, our neural face renderer may fail and produce artifacts, see Figure 13(b).

6 Conclusion

We have presented a novel real-time approach for synthesizing photorealistic video portraits from an input audio and a target video. We proposed an Audio2Expression network to predict the facial expression parameters for a target actor from any speech audio. By blending the predicted facial expression parameters and reconstructed 3D face parameters from the target video, synthetic face images are rendered in sync with the audio. Finally, we train a neural face renderer with an elegant mask expansion strategy that translates synthetic renderings into photorealistic video portraits.

Qualitative and quantitative evaluations show that our method outperforms previous general-purpose audio-driven video portrait approaches, except for Audio2Obama [46], a specifically tailored method that only works with extensive, consistent audio and video taken from the same actor. The user study confirmed that our results are compelling and generally preferred to other general-purpose audio-driven methods.

Our proposed method provides benefits for several VR/AR applications, including photorealistic virtual news anchors, and virtual education and training. It also supports a large variety of applications, such as online digital voice assistant enhancement and video conferencing, especially when the network bandwidth is limited. We believe our approach takes an important step towards solving this challenging task and it could potentially be combined with even more VR/AR applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Project Number: 61902012 and 61932003), RCUK grant CAMERA (EP/M023281/1), and an EPSRC-UKRI Innovation Fellowship (EP/S001050/1).

References

[1] H. Averbuch-Elor, D. Cohen-Or, J. Kopf, and M. F. Cohen. Bringing portraits to life. *ACM Trans. Graph.*, 36(6):196:1–13, 2017. doi: 10.1145/3130800.3130818

[2] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh. Recycle-GAN: Unsupervised video retargeting. In *ECCV*, 2018. doi: 10.1007/978-3-030-01228-1_8

[3] F. Berthouzoz, W. Li, and M. Agrawala. Tools for placing cuts and transitions in interview video. *ACM Trans. Graph.*, 31(4), 2012.

[4] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, pp. 187–194, 1999. doi: 10.1145/311535.311556

[5] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *ICCV*, 2017. doi: 10.1109/ICCV.2017.116

[6] C. Cao, D. Bradley, K. Zhou, and T. Beeler. Real-time high-fidelity facial performance capture. *ACM Trans. Graph.*, 34(4):46:1–9, 2015. doi: 10.1145/2766943

[7] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. doi: 10.1109/TVCG.2013.249

[8] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244

[9] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Face & Gesture*, 2018. doi: 10.1109/FG.2018.00020

[10] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros. Everybody dance now. In *ICCV*, pp. 5933–5942, 2019. doi: 10.1109/ICCV.2019.00603

[11] L. Chen, R. K. Maddox, Z. Duan, and C. Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *CVPR*, pp. 7824–7833, 2019. doi: 10.1109/CVPR.2019.00802

[12] J. S. Chung, A. Jamaludin, and A. Zisserman. You said that? In *BMVC*, 2017.

[13] J. S. Chung and A. Zisserman. Lip reading in the wild. In *ACCV*, 2016.

[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006. doi: 10.1121/1.2229005

[15] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black. Capture, learning, and synthesis of 3D speaking styles. In *CVPR*, 2019. doi: 10.1109/CVPR.2019.01034

[16] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *CVPR Workshops*, 2019. doi: 10.1109/CVPRW.2019.00038

[17] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter. 3D morphable face models—Past, present, and future. *ACM Trans. Graph.*, 39(5):157:1–38, 2020. doi: 10.1145/3395208

[18] O. Fried, A. Tewari, M. Zollhöfer, A. Finkelstein, E. Shechtman, D. B. Goldman, K. Genova, Z. Jin, C. Theobalt, and M. Agrawala. Text-based editing of talking-head video. *ACM Trans. Graph.*, 38(4):68:1–14, 2019. doi: 10.1145/3306346.3323028

[19] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt. Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. Graph.*, 35(3):28:1–15, 2016. doi: 10.1145/2890493

[20] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman. Unsupervised training for 3D morphable model regression. In *CVPR*, pp. 8377–8386, 2018. doi: 10.1109/CVPR.2018.00874

[21] M. Gonzalez-Franco, A. Steed, S. Hoogendyk, and E. Ofek. Using facial animation to increase the enfacement illusion and avatar self-identification. *IEEE Transactions on Visualization and Computer Graphics*, 26(5):2023–2029, 2020. doi: 10.1109/TVCG.2020.2973075

[22] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

[23] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6):1294–1307, 2019. doi: 10.1109/TPAMI.2018.2837742

[24] N. Harte and E. Gillen. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. doi: 10.1109/TMM.2015.2407694

[25] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90

[26] C. Heeter. Being there: The subjective experience of presence. *Presence: Teleoperators & Virtual Environments*, 1(2):262–271, 1992.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pp. 5967–5976, 2017. doi: 10.1109/CVPR.2017.632

[28] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–12, 2017. doi: 10.1145/3072959.3073658

[29] H. Kim, M. Elgharib, M. Zollhöfer, H.-P. Seidel, T. Beeler, C. Richardt, and C. Theobalt. Neural style-preserving visual dubbing. *ACM Trans. Graph.*, 38(6):178:1–13, 2019. doi: 10.1145/3355089.3356500

[30] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Trans. Graph.*, 37(4):163:1–14, 2018. doi: 10.1145/3197517.3201283

[31] H. Kim, M. Zollhöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt. InverseFaceNet: Deep monocular inverse face rendering. In *CVPR*, pp. 4625–4634, 2018. doi: 10.1109/CVPR.2018.00486

[32] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[33] L. Liu, W. Xu, M. Zollhöfer, H. Kim, F. Bernard, M. Habermann, W. Wang, and C. Theobalt. Neural animation and reenactment of human actor videos. *ACM Trans. Graph.*, 38(5):139:1–14, 2019. doi: 10.1145/3333002

[34] M. Meshry, D. B. Goldman, S. Khamis, H. Hoppe, R. Pandey, N. Snavely, and R. Martin-Brualla. Neural rerendering in the wild. In *CVPR*, 2019. doi: 10.1109/CVPR.2019.00704

[35] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv:1411.1784, 2014.

[36] M. Murcia-López, T. Collingwoode-Williams, W. Steptoe, R. Schwartz, T. J. Loving, and M. Slater. Evaluating virtual reality experiences through participant choices. In *IEEE VR*, pp. 747–755, 2020. doi: 10.1109/VR46266.2020.00098

[37] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. WaveNet: A generative model for raw audio. arXiv:1609.03499, 2016.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

[39] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal based Surveillance*, pp. 296–301, 2009.

[40] K. R. Prajwal, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. V. Jawahar. Towards automatic face-to-face translation. In *Proceedings of the International Conference on Multimedia*, pp. 1428–1436, 2019. doi: 10.1145/3343031.3351066

[41] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *SIGGRAPH*, pp. 497–500, 2001. doi: 10.1145/383259.383317

[42] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, pp. 5553–5562, 2017. doi: 10.1109/CVPR.2017.589

[43] J. Roth, Y. T. Tong, and X. Liu. Adaptive 3D face reconstruction from unconstrained photo collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2127–2141, 2017. doi: 10.1109/TPAMI.2016.2636829

[44] F. Shi, H.-T. Wu, X. Tong, and J. Chai. Automatic acquisition of high-fidelity facial performances using monocular videos. *ACM Trans. Graph.*, 33(6):222:1–13, 2014. doi: 10.1145/2661229.2661290

[45] L. Song, W. Wu, C. Qian, R. He, and C. C. Loy. Everybody's talkin': Let me talk as you want. arXiv:2001.05201, 2020.

[46] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–13, 2017. doi: 10.1145/3072959.3073640

[47] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM Trans. on Graph.*, 36(4), July 2017. doi: 10.1145/3072959.3073699

[48] A. Tewari, O. Fried, J. Thies, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B. Goldman, and M. Zollhöfer. State of the art on neural rendering. *Comput. Graph. Forum*, 39(2):701–727, 2020. doi: 10.1111/cgf.14022

[49] A. Tewari, M. Zollhöfer, F. Bernard, P. Garrido, H. Kim, P. Pérez, and C. Theobalt. High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):357–370, 2020. doi: 10.1109/TPAMI.2018.2876842

[50] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV*, 2020.

[51] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4):66:1–12, 2019. doi: 10.1145/3306346.3323035

[52] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. *Communications of the ACM*, 62(1):96–104, 2018. doi: 10.1145/3292039

[53] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner. Image-guided neural object rendering. In *ICLR*, 2020.

[54] K. Vougioukas, S. Petridis, and M. Pantic. Realistic speech-driven facial animation with GANs. *IJCV*, 128:1398–1413, 2020. doi: 10.1007/s11263-019-01251-8

[55] M. Wang, X.-Q. Lyu, Y.-J. Li, and F.-L. Zhang. VR content creation and exploration with deep learning: A survey. *Computational Visual Media*, 6:3–28, 2020. doi: 10.1007/s41095-020-0162-z

[56] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro. Few-shot video-to-video synthesis. In *NeurIPS*, 2019.

[57] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018.

[58] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861

[59] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with learning-based personalized head pose. arXiv:2002.10137, 2020.

[60] Z. Yi, H. Zhang, P. Tan, and M. Gong. DualGAN: Unsupervised dual learning for image-to-image translation. In *ICCV*, pp. 2868–2876, 2017. doi: 10.1109/ICCV.2017.310

[61] L. Yu, J. Yu, M. Li, and Q. Ling. Multimodal inputs driven talking face generation with spatial-temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. doi: 10.1109/TCSVT.2020.2973374

[62] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *ICCV*, 2019. doi: 10.1109/ICCV.2019.00955

[63] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI*, 2019. doi: 10.1609/aaai.v33i01.33019299

[64] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pp. 2242–2251, 2017. doi: 10.1109/ICCV.2017.244

[65] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. State of the art on monocular 3D face reconstruction, tracking, and applications. *Comput. Graph. Forum*, 37(2):523–550, 2018. doi: 10.1111/cgf.13382