International Journal of Computer Vision (2024) 132:2713–2728
https://doi.org/10.1007/s11263-024-02007-9

ReliTalk: Relightable Talking Portrait Generation from a Single Video

Haonan Qiu[1] · Zhaoxi Chen[1] · Yuming Jiang[1] · Hang Zhou[2] · Xiangyu Fan[3] · Lei Yang[3] · Wayne Wu[3] · Ziwei Liu[1]

Received: 1 August 2023 / Accepted: 14 January 2024 / Published online: 16 February 2024
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract
Recent years have witnessed great progress in creating vivid audio-driven portraits from monocular videos. However, how to seamlessly adapt the created video avatars to other scenarios with different backgrounds and lighting conditions remains unsolved. On the other hand, existing relighting studies mostly rely on dynamically lighted or multi-view data, which are too expensive for creating video portraits. To bridge this gap, we propose **ReliTalk**, a novel framework for relightable audio-driven talking portrait generation from monocular videos. Our key insight is to decompose the portrait's reflectance from implicitly learned audio-driven facial normals and images. Specifically, we involve 3D facial priors derived from audio features to predict delicate normal maps through implicit functions. These initially predicted normals then take a crucial part in reflectance decomposition by dynamically estimating the lighting condition of the given video. Moreover, the stereoscopic face representation is refined using the identity-consistent loss under simulated multiple lighting conditions, addressing the ill-posed problem caused by limited views available from a single monocular video. Extensive experiments validate the superiority of our proposed framework on both real and synthetic datasets. Our code is released in (https://github.com/arthur-qiu/ReliTalk).

Keywords Relighting · Talking face · Portrait Generation · Relightable Portrait

1 Introduction

Creating personalized audio-driven talking portraits has many applications in teleconferencing, video production, VR/AR games, and the movie industry. Given its great potential, research on talking face generation (Taylor et al., 2017; Thies et al., 2020; Zhou et al., 2019b, 2021; Zhang et al., 2021c; Ji et al., 2021) has enjoyed massive popularity in recent years, with emphasis on creating lip-synced (Prajwal et al., 2020; Thies et al., 2020) portraits with diverse head motions, talking styles, and emotions (Yi et al., 2020; Wu et al., 2021). However, the ability to change the lighting conditions of audio-driven portraits is still under-explored, which is critical to real-world applications as we expect the portrait in the foreground to be seamlessly harmonized with backgrounds under different illuminations.

To generate a relightable talking portrait from a single video, we argue that the underlying model should be capable of (1) estimating fine-grained 3D head geometry from monocular videos, (2) reflectance decomposition without any extra annotations, and (3) generalizing to driven audios. However, most learning-based methods either operate only on the 2D plane (Zhou et al., 2019b, 2021; Prajwal et al., 2020), or leverage structural intermediate representations

Communicated by Gang Hua.

✉ Ziwei Liu
ziwei.liu@ntu.edu.sg

Haonan Qiu
HAONAN002@e.ntu.edu.sg

Zhaoxi Chen
ZHAOXI001@e.ntu.edu.sg

Yuming Jiang
YUMING002@e.ntu.edu.sg

Hang Zhou
zhouhang@link.cuhk.edu.hk

Xiangyu Fan
fanxy1993@gmail.com

Lei Yang
yanglei@sensetime.com

Wayne Wu
wuwenyan0503@gmail.com

[1] S-Lab, Nanyang Technological University, Singapore, Singapore

[2] The Chinese University of Hong Kong, Shatin, Hong Kong

[3] SenseTime Research, Shenzhen, China

Fig. 1 Relighting Talking Portrait with Assigned Background. (Left) Our method takes a monocular video as input and estimates the corresponding normal and albedo which can be driven by audio. (Right) Talking portrait renderings with different illuminations, where lighting and shading are placed at the bottom. The rightmost three are relighted by HDR background images. Only a single video is required as the training data, without any extra annotations

(Chen et al., 2019; Cudeiro et al., 2019a; Ji et al., 2021; Thies et al., 2020; Wu et al., 2021; Zhou et al., 2020), and neural radiance fields (Guo et al., 2021; Yao et al., 2022; Shen et al., 2022). No fine-grained 3D geometry can be acquired for reflectance decomposition in these studies. On the other hand, adapting existing relighting techniques (Sun et al., 2019; Wang et al., 2020; Pandey et al., 2021; Zhang et al., 2021a) is too expensive for audio-driven video portraits given their dependence on multi-view or dynamically lighted data.

To bridge this gap, we propose **ReliTalk**, a novel framework for relightable audio-driven talking portrait generation that only requires a single monocular video as input, as shown in Fig. 1. Our key insight is the self-supervised implicit decomposition of geometry and reflectance, both of which can be further driven by input audios. In specific, the proposed approach first extracts expression- and pose-related representations based on 3D facial priors (Li et al., 2017), and refines them into delicate normal maps through implicit functions. The initial normals then take a critical role in reflectance decomposition, which disentangles the human head as a set of intrinsic normal, albedo, diffuse and specular maps, by dynamically estimating the lighting condition of the given video. To get rid of leveraging knowledge from expensive capturing data (*i.e.* Light Stage (Debevec et al., 2000)), we carefully design several learning objectives to decompose the human portrait into corresponding maps from monocular videos, which will be introduced in the following sections.

To learn the audio-to-face mapping that better generalizes to unseen audio, we introduce mesh-aware guidance to assist the lip-syncing especially when the training video is too short to cover enough audio variance. Specifically, we use a model pre-trained on the VOCA dataset (Cudeiro et al., 2019b) to obtain lip-related meshes as the additional guidance. Phoneme-related features and lip-related meshes are separately encoded and then concatenated to achieve more accurate audio-driven animations. Phoneme-related features enable the network to learn richer mouth shapes and mesh-aware features provide coarse information on the opening and closing of lips, even if the input audio is far away from the audio used in training.

Natural talking portrait videos usually provide a limited perspective of the target persons when they face the camera without turning around. Plus, the lack of multi-view information inherently negatively impacts an accurate estimation of 3D geometry. To address the ill-posed inverse problem of geometry and reflectance decomposition caused by single-view, limited motion variance, and unknown illuminations, we design identity-consistent supervision (ICS) with simulated multiple lighting conditions to refine normal maps. The key insight is that we relight the human portrait on-the-fly during the training stage, by sampling different lights and using identity-consistent loss to update normal maps.

We evaluate our approach on both real and synthetic datasets. Overall, **ReliTalk** drives and relights dynamic human portraits in high fidelity, outperforming other methods on both perceptual quality and reconstruction correctness. Our contributions are summarized as follows:

- We propose a novel framework **ReliTalk** that learns relightable audio-driven talking portrait generation and only requires a single monocular portrait video.
- We propose the additional audio-to-mesh guidance to improve the mapping accuracy especially when the single training video only has a limited audio variance.
- We design identity-consistent supervision with simulated multiple lighting conditions, addressing the ill-pose problem caused by limited views available from the single video.

2 Related Work

Inverse Rendering Recovering and disentangling the appearance of observed images into geometry and reflectance is a long-standing problem in the field of computer vision and graphics. Prior works (Barron & Malik, 2014; Liu et al., 2019) address this challenge by physical-based priors on synthetic image data. However, they fail to extract the underlying 3D representation. Later approaches (Chan et al., 2022; Or-El et al., 2022; Xu et al., 2022; Sun et al., 2022; Zhao et al., 2022; Pan et al., 2020; Chan et al., 2021) successfully extract the 3D representations by the 3D generator and refine the output using image-based CNN networks. Recently, methods based on implicit representation (Zhang et al., 2021b; Srinivasan et al., 2021) propose learning 3D reflectance and geometry from multi-view images. In this work, we aim to tackle a harder problem, *i.e.*, inverse rendering from a monocular video of a talking human face. Note that, limited-view information can be accessed as the person is always oriented toward the front.

Portrait Relighting One-Light-at-A-Time (OLAT) capturing system allows for obtaining detailed portrait geometry and reflectance. Many methods based on it have achieved impressive success (Sun et al., 2019; Wang et al., 2020; Pandey et al., 2021; Zhang et al., 2021a). However, it is only applicable in a constrained environment due to its complexity and expense. Other methods (Zhou et al., 2019a; Hou et al., 2021, 2022; Caselles et al., 2023) simulate some multi-lighting data and train the network to predict relighted results. Due to their limited simulation methods, the final results are far away from OLAT-based methods. Yeh et al. (2022) synthesizes a high-quality multi-lighting dataset but it is still not available to the public. Another simplified strategy that requires the user to capture a selfie video or a sequence of images to gain multi-view information is proposed (Nestmeyer et al., 2020; Wang et al., 2022). And Relighting4D (Chen & Liu, 2022) can even relight dynamic humans with free viewpoints only from videos. However, their rendering quality is totally tied to the accuracy of geometry, requiring enough viewpoints from videos. Our method is able to relight portraits with finer details from the monocular portrait video even without much multi-view information available.

Audio-driven Talking Face Face animation has wild applications, drawing great research interest in computer vision and graphics. Recent methods for audio-driven animation (Cudeiro et al., 2019a; Fan et al., 2022; Karras et al., 2017; Richard et al., 2021; Suwajanakorn et al., 2017; Kim et al., 2018) are usually data-driven and can be divided into two categories. One is generalized animation (Cudeiro et al., 2019a; Fan et al., 2022; Richard et al., 2021), which utilizes some large datasets which contain the pair data of audio/speech to lip/face. Wave2Lip (Prajwal et al., 2020) trains a mapping from audio to lips on LRS2 (Chung et al., 2017). Instead of learning a highly heterogeneous and non-linear mapping from audio to video directly, Everybody's Talkin (Song et al., 2022) additionally involves the statistical linear 3D face model and builds an easier map from audio to parameters of 3DMM (Blanz & Vetter, 1999). Our proposed method takes a similar strategy that drives the whole portrait through controlling the parameters of FLAME model (Li et al., 2017). The other one is personalized animation (Suwajanakorn et al., 2017; Karras et al., 2017; Tang et al., 2022), which usually does not rely on a large dataset for training and only builds one model for each person. Recently, with the emergence of Neural Radiance Fields (NeRF) (Mildenhall et al., 2020; Barron et al., 2021b, a), many NeRF-based audio-driven methods are proposed (Guo et al., 2021; Liu et al., 2022; Yao et al., 2022). However, those methods can not drive the portraits well when meeting novel audio. DFRF (Shen et al., 2022) improves this issue with a pre-trained base model but the final results are still not satisfactory.

3 Our Approach

Given a monocular video of a talking portrait, our framework can re-render the human portrait with novel illuminations driven by the input audio. Denote the input video with unknown illuminations as $\mathbf{V} = \{I_1, I_2, ..., I_t\}$ with audio sequence $\mathbf{a} = \{a_1, a_2, ..., a_t\}$, where t is the number of frames. The key aim of our framework is to extract the geometry and reflectance information from video $\mathbf{V}$ in an unsupervised manner, and the geometry deformation is driven by the audio accordingly. Specifically, we neurally model the expression- and pose-related geometry of human heads based on the FLAME model (Li et al., 2017). Then, an audio-to-geometry mapping is learned to drive the portrait and also provide a good initial normal estimation (Sect. 3.1). Meanwhile, the reflectance components, *i.e.,* normal N, albedo A, shading S_{shad}, and specular S_{spec} maps, are decomposed via carefully designed priors (Sect. 3.2). During training, the lighting condition L of the given video is estimated on-the-fly, and the training objective is reconstructing the whole video. In addition, multiple lighting conditions are randomly simulated for identity-consistent supervision which further refines geometry estimation. With the well-disentangled geometry and reflectance, we use audio from the user to drive the portrait by controlling expression and pose coefficients, then render it with any desired illuminations, which seamlessly harmonizes with the background. The whole pipeline is shown in Fig. 2.

In this paper, $I, N, A \in \mathbb{R}^{3\times H\times W}$, $S_{shad}, S_{spec} \in \mathbb{R}^{1\times H\times W}$ where H and W are height and width respectively.

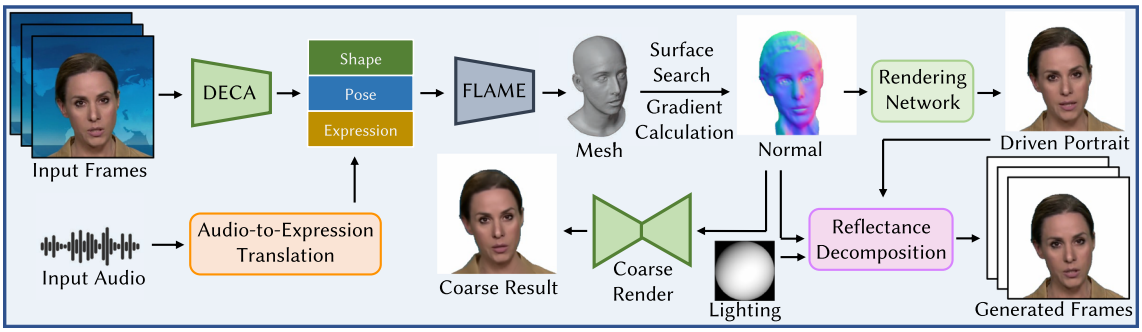

Fig. 2 Overview of Our Proposed Framework. Denote the input video with unknown illuminations as $\mathbf{V} = \{I_1, I_2, ..., I_t\}$ with audio sequence $\mathbf{a} = \{a_1, a_2, ..., a_t\}$, where t is the number of frames. Generally, our aim is to extract the geometry and reflectance information from video $\mathbf{V}$ in an unsupervised manner then drive the geometry deformation according to the audio

3.1 Audio-Driven Synthesis

Expression- and Pose-related Geometry Estimating the surface normal of talking portraits from monocular videos is a non-trivial task, given the ill-posed nature of single-view reconstruction. To address this issue, we leverage a parametric model, FLAME (Li et al., 2017), as the human head prior to modeling the expression- and pose-related human portrait:

$$\text{FLAME}(\beta, \theta, \psi) : \mathbb{R}^{|\beta|\times|\theta|\times|\psi|} \rightarrow \mathbb{R}^{n\times 3}, \tag{1}$$

which takes coefficients of shape $\beta \in \mathbb{R}^{|\beta|}$, pose $\theta \in \mathbb{R}^{|\theta|}$, and expression $\psi \in \mathbb{R}^{|\psi|}$ as input. We use the off-the-shelf tool (Feng et al., 2021) to estimate those parameters. Intuitively, this parametric human portrait model offers a good initialization of the 3D geometry, which facilitates further refinement.

However, this initial parametric portrait model is not well-aligned with the details of the given human portrait. To refine the initial model, we use nearest surface intersection search (Zheng et al., 2022) to optimize the initial mesh and calculate the normal N as the normalized gradient on the surface. This normal N will be further optimized during the reflectance decomposition process (Sect. 3.2).

Mesh-Aware Audio-to-Expression Translation From the perspective of the mapping function, learning a direct mapping from audio to talking video is hard due to its high-dimensional property. In contrast, mapping audio signals to expressions and poses of the head is much easier. To enable robust talking portrait generation, we leverage a mesh-aware audio-to-expression translation strategy. Benefiting from FLAME (Li et al., 2017) based design, our extracted head geometry is expression- and pose-related. We first use DeepSpeech (Amodei et al., 2016) to extract phoneme-related audio features $f_{\text{pho}} \in \mathbb{R}^{16\times 29}$:

$$f_{\text{pho}} = \text{DeepSpeech}(a). \tag{2}$$

Then extracted audio features f_{pho} are fed to a model pre-trained on the VOCA dataset (Cudeiro et al., 2019b) to predict lip-related mesh vertices $V_{\text{lip}} \in \mathbb{R}^{N_V\times 3}$ (N_V is the selected vertex number) as the additional guidance:

$$V_{\text{lip}} = F_{\text{mesh}}(V_{\text{template}}, f_{\text{pho}}), \tag{3}$$

where V_{template} is the zero-pose template for audio features.

Lip-related vertices and phoneme-related features are separately encoded and concatenated to predict expressions and poses of the FLAME model by a learnable network:

$$\hat{\psi}, \hat{\theta} = F_{\text{exp}}(E_{\text{lip}}(V_{\text{lip}}), E_{\text{pho}}(f_{\text{pho}})), \tag{4}$$

where E_{lip} and E_{pho} are two feature encoders and F_{exp} is a network that concatenates two kinds of features and predicts expressions and poses. Meanwhile, to address the unstable prediction caused by a single audio frame, we input neighboring frames and use attention layers in F_{exp} to integrate multi-frame audio information. This learning process is supervised by $\mathcal{L}_{\text{exp}} = \|\hat{\psi} - \psi\|_2^2 + \|\hat{\theta} - \theta\|_2^2$.

Neural Video Rendering Network Gaining the new driven coefficients ψ, θ, we can send them to Eq. 1 and recalculate the geometry to get a new normal $\hat{N}$ that fits input audio. Here we find that it is non-trivial to faithfully relate the audio signals and all face deformations (*e.g.*let@tokeneonedothead movement). The translation network F_{exp} will perform poorly if required to fit all poses. Therefore, we only predict lip-related poses and directly use the sequence of lip-unrelated poses from existing videos. In this way, we only need to regenerate lip-related areas (including cheek and chin) and blend lip-unrelated areas from the existing videos.

We first use a ResNet based local network F_{local} to translate the newly generate normal to lip-related areas. Using eroded lip-unrelated areas from existing videos as background and the output of the first network, another blending

network F_{blend} is used to output the blended image $\hat{I}$:

$$\hat{I} = F_{\text{blend}}(F_{\text{local}}(\hat{N}) \odot M, I \odot (1 - M^d)), \tag{5}$$

where M is the lip-related area and M^d is the dilated area for the network F_{blend} to inpaint. This process is learned by:

$$\mathcal{L}_{\text{rgb}}^{\text{local}} = \|F_{\text{local}}(\hat{N}) \odot M - I \odot M\|_2^2, \tag{6}$$

$$\mathcal{L}_{\text{rgb}}^{\text{blend}} = \|\hat{I} - I\|_2^2, \tag{7}$$

$$\mathcal{L}_{\text{per}}^{\text{blend}} = \|\text{VGG}(\hat{I}) - \text{VGG}(I)\|_2^2, \tag{8}$$

where $\mathcal{L}_{\text{per}}^{\text{blend}}$ is the perceptual loss and VGG represents a pretrained face VGG network (Parkhi et al., 2015) and returns extracted embedding features. It is only added to the blending network to generate vivid results while the local network is only supposed to generate the rough lip area.

3.2 Reflectance Decomposition

To enable rendering the talking portrait under novel illuminations, the reflectance and environmental lighting should be appropriately disentangled and estimated.

Lighting Following previous work (Ramamoorthi & Hanrahan, 2001; Barron & Malik, 2014; Basri & Jacobs, 2003; Wang et al., 2008; Shu et al., 2017; Zhou et al., 2019a; Hou et al., 2021), the environmental lighting $L \in \mathbb{R}^9$ is represented as a 9-dimensional spherical harmonics coefficient vector. However, the lighting conditions of online talking videos are unknown, which makes it hard for inverse rendering. Inspired by Relighting4D (Chen & Liu, 2022), we first initialize the lighting L from the front of the human face and then treat it as a trainable parameter to optimize. During the inference, given HDR lighting will be converted to a 9-dimensional spherical harmonics coefficient vector.

Normal Map Although $\hat{N}$ provides a rough estimation of portrait geometry, its deviation from the real geometry will be amplified in the relighting. To further refine the geometry while still keeping the structure of $\hat{N}$, we use a network F_{normal} to predict normal residual:

$$\delta N = F_{\text{normal}}(\hat{I}, \hat{N}). \tag{9}$$

We add an L_1 regularization on δN, i.e., $\mathcal{L}_{\delta N} = \|\delta N\|_1$. The final predicted normal N is $N = \hat{N} + \delta N$.

Shading Map Given the normal N and lighting L, we can calculate the shading map S_{shad} using a network F_{shad} conditioned on the normal and lighting:

$$S_{\text{shad}} = F_{\text{shad}}(N, L), \tag{10}$$

Albedo Map To represent the illumination-invariant base color of the human face, we use a network F_{albedo} to predict the albedo map A from the appearance:

$$A = F_{\text{albedo}}(\hat{I}). \tag{11}$$

Although there is no ground truth for albedo in our setting, it is supposed to have two physical priors: smoothness and parsimony (Barron & Malik, 2014). Smoothness requires that variation in the albedo map tends to be small and sparse. To achieve that, we use total variation regularization on the skin area:

$$\begin{aligned}\mathcal{L}_{\text{smooth}}(A) = &\sum_{h=1}^{H}\sum_{w=1}^{W} \left\|\boldsymbol{\beta}_{h+1,w} - \boldsymbol{\beta}_{h,w}\right\|_2^2 \\ &+ \sum_{h=1}^{H}\sum_{w=1}^{W} \left\|\boldsymbol{\beta}_{h,w+1} - \boldsymbol{\beta}_{h,w}\right\|_2^2,\end{aligned} \tag{12}$$

where $\boldsymbol{\beta}_*$ are the values of albedo A within the skin area.

In addition to piece-wise smoothness, the second property we expect from albedo map is parsimony, which means that the palette with which an albedo image was painted should be small. This property holds only when it is a soft constraint to make the palette sparse enough. As for the parsimony prior, we penalize the network by minimizing the entropy of the albedo map (Chen & Liu, 2022):

$$\mathcal{L}_{\text{parsimony}} = \mathbb{E}[-\log(p(A))], \tag{13}$$

where $p(\cdot)$ is the probability density function (PDF). To address the difficulty of estimating the PDF of the continuous variable albedo map A during training, we use Monte Carlo sampling to obtain a soft approximation of a Gaussian histogram at predefined bins for estimating the PDF of A.

Specular Map Prior works (He et al., 2016; Shu et al., 2017) for portrait relighting, especially which require no One-Light-at-A-Time (OLAT) data, only employ simple diffuse lighting to model the human face. However, given the fact that specular phenomenons widely appear on human faces, it is key to the photorealistic rendering to model the specular effects. Therefore, we leverage Blinn–Phong model (1977) to incorporate specular component as:

$$R_{\text{spec}}(N, \omega_i, \omega_o) = \frac{s+2}{2\pi}\left(h(\omega_i, \omega_o) \cdot N\right)^s, \tag{14}$$

where $h(\omega_i, \omega_o) = \text{normalize}(\omega_i + \omega_o)$, and s is the Phong exponent that controls the apparent smoothness of the surface. Then the specular map S_{spec} can be calculated by the accumulation of $R^{\text{spec}}(N, \omega_i, \omega_o)$ under illumination from different directions:

$$S_{\text{spec}} = F_{\text{spec}}(N, L) = \sum_{\omega_i}(L(\omega_i) \odot R_{\text{spec}}(N, \omega_i, \omega_o)), \tag{15}$$

in which ω_o is always towards the front in this paper. In experiments, we also find that the specular produced by Blinn-Phong model can never perfectly align with the real face in the video. Inspired by SunStage (Wang et al., 2022), we use another network F_{cspec} to predict a coefficient map $C_{\text{spec}} \in \mathbb{R}^{1\times H\times W}$ for flexibly adjusting the final specular:

$$C_{\text{spec}} = F_{\text{cspec}}(\hat{I}, N). \tag{16}$$

For the coefficient map, we apply a TV loss mentioned in Eq. 12 to avoid checkerboard artifacts. Finally, we synthesize the video frame $\tilde{I}$ via image-based rendering:

$$F_{\text{render}} : \tilde{I} = A \odot (S_{\text{shad}} + C_{\text{spec}} \odot S_{\text{spec}}), \tag{17}$$

where $\odot$ denotes the element-wise product. And the training objective is the reconstruction loss against input frames:

$$\mathcal{L}_{\text{rgb}}^{\text{render}} = \|\tilde{I} - I\|_2^2. \tag{18}$$

Identity-Consistent Supervision with Relighting Coarse renderer F_{coarse} synthesizes RGB pixel values according to the normal N. Without multi-view supervision, the face area in highlights will be regarded as raised part even if it is smooth originally, leading to artifacts during the relighting. To address this issue, we propose identity-consistent supervision with simulated multiple lighting conditions, which is performed on-the-fly during training. We assume that a well-trained face recognition network can extract similar embedding when the lighting condition varies. Therefore, after the decomposition is nearly converged, we randomly sample a new lighting condition and reinforce the identity consistency between the two rendered images with different lighting conditions:

$$\mathcal{L}_{\text{consistent}} = \|E_{\text{id}}(I^{\text{relight}}) - E_{\text{id}}(I)\|_2^2, \tag{19}$$

where I^{relight} is the rendering under the randomly sampled lighting, and E_{id} is the embedding extracted by a pre-trained face recognition network (Schroff et al., 2015).

3.3 Optimization and Inference

During the training phase, training the entire framework directly may cause the networks to learn the locally optimized results of each decomposed map, as there are no ground truths available for each component of reflectance decomposition. Therefore, we first train networks for audio-driven synthesis to learn a rough expression- and pose-related geometry.

Yet, there is no off-the-shelf ground truth for normal map N to supervise geometry refinement. To address it, a coarse renderer F_{coarse} is used to predict the RGB result $\hat{I}$ conditioning on normal N, which is supervised by image reconstruction loss. This self-supervised learning process encourages the normal map to obtain more details of surface shape, without requiring extra annotations. After a rough normal map N is gained, it is combined with an RGB portrait image as inputs of reflectance decomposition for further optimization and also stabilize the decomposition.

The overall loss is:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\text{rgb}}^{\text{local}}\mathcal{L}_{\text{rgb}}^{\text{local}} + \lambda_{\text{rgb}}^{\text{blend}}\mathcal{L}_{\text{rgb}}^{\text{blend}} + \lambda_{\text{per}}^{\text{blend}}\mathcal{L}_{\text{per}}^{\text{blend}} \\ & + \lambda_{\text{rgb}}^{\text{render}}\mathcal{L}_{\text{rgb}}^{\text{render}} + \lambda_{\delta N}\mathcal{L}_{\delta N} + \lambda_{\text{consistent}}\mathcal{L}_{\text{consistent}} \\ & + \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \lambda_{\text{parsimony}}\mathcal{L}_{\text{parsimony}} + \lambda_{\text{smooth}}^{\text{total}}\mathcal{L}_{\text{smooth}}^{\text{total}},\end{aligned} \tag{20}$$

where λ's are the weights and are set to 1, 1, 100, 1, 1, 3, 1, 0.001, and 1 respectively. Here $\mathcal{L}_{\text{smooth}}^{\text{total}}$ is similar to Eq. 12 but is added to all decomposed maps:

$$\begin{aligned}\mathcal{L}_{\text{smooth}}^{\text{total}} = {} & \mathcal{L}_{\text{smooth}}(A) + \mathcal{L}_{\text{smooth}}(N) \\ & + \mathcal{L}_{\text{smooth}}(R_{\text{spec}}) + \mathcal{L}_{\text{smooth}}(C_{\text{spec}}).\end{aligned} \tag{21}$$

In the inference phase, new audio will drive the portrait by controlling expression and pose coefficients. Meanwhile, desired illuminations will replace the learned light L of the original video to relight the whole video, thus seamlessly harmonizing with the background.

4 Experiments

4.1 Implementation Details

Network Architecture In audio-driven synthesis, lip-related feature encoder E_{exp} and phoneme-related feature encoder E_{pho} both use 1D convolutional neural networks. Decoder F_{exp} is also a 1D convolutional neural network but with the self-attention mechanism (Zhang et al., 2019) to predict pose and expression coefficients through 8 adjacent frames. For Local network F_{local}, we use the ResNet (He et al., 2016) with 6 residual blocks. While for blending network F_{blend}, we use U-Net of depth 5 with dilated convolutions (Thies et al., 2020). To gain coherent results, we adjust the mask size to leave some missing area between the generated lip area and the given background area, which will be inpainted by F_{blend}. As shown in Fig. 3, we choose the area with 80×80 resolution around the mouth as the Driven Area and remove the area with 120×120 resolution around the mouth as the Existing Area. Here we also add the lip area generated by Wave2Lip (Prajwal et al., 2020) as the additional input of F_{blend} to increase the performance when the input audio is far away from the audio used in training (i.e. audio from a new person).

For reflectance decomposition, we choose U-Net (Isola et al., 2017) of depth 8 as the architecture of specular weight

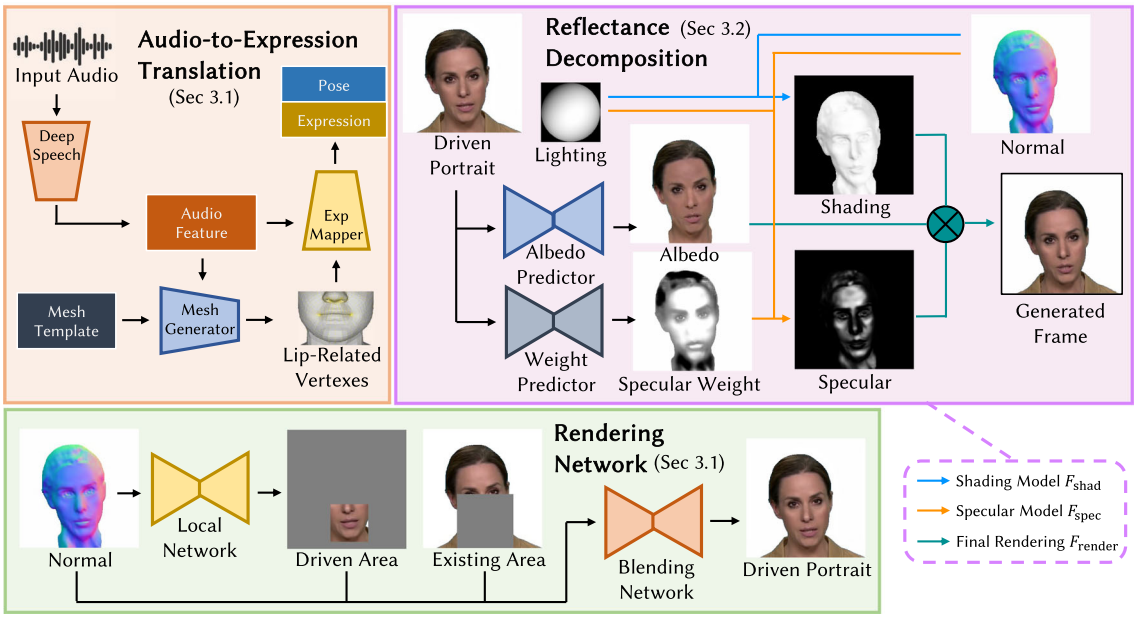

Fig. 3 Details of Our Proposed Framework. Our framework decomposes the video I into a set of normal N, albedo A, shading S_{shad}, and specular S_{spec} maps. Specifically, we neurally model the expression- and pose-related geometry of human heads based on the FLAME model (Li et al., 2017). Then, the reflectance components are decomposed via multiple carefully designed priors (Sect. 3.2). With the well-disentangled geometry and reflectance, we use audio from the user to drive the human portrait by controlling expression and pose coefficients, then render it with any desired illuminations, which seamlessly harmonizes with the background

predicter F_{cspec}. But to gain smoother albedo maps and normal residuals, we choose ResNet (He et al., 2016) with 6 residual blocks as the architecture of albedo predictor F_{albedo} and normal residual predictor F_{normal}.

Running Time We conduct our experiments on a single GPU of NVIDIA V100. Each person will take around 1 day for training and the inference time is around 0.6 s per frame.

4.2 Dataset

Real Video Data AD-NeRF (Guo et al., 2021) and HDTF (Zhang et al., 2021c) collect several high-resolution talking videos in different scenes to better evaluate the generation performance. Following this practice, we choose celebrity videos whose protagonists are news anchors, entrepreneurs, or presidents from YouTube as our real video set. We collect 8 public videos with an average length of 3 min from 7 identities. We split each video with around 80% frames for training and 20% frames for evaluation. These videos are all available online and we will provide corresponding source links for reproduction purposes.

Synthetic Video Data Talking videos with ground-truth illuminations are not available from online collections. To evaluate our relighting algorithm quantitatively, we synthesize some talking videos with the same motion sequence but different lighting conditions within the modern graphic pipeline, as shown in Fig. 4. Specifically, we render 6 sequences (2 min, 25 fps), for each person in 10 different lighting conditions with Cycles renderer (Hess, 2013) in Blender (Community, 2018), a photorealistic ray-tracing renderer. All mesh models, textures, and displacement maps are released by FaceScape (Yang et al., 2020; Zhu et al., 2021). We combine displacement maps and textures in a physically-based skin material featuring sub-surface scattering (Christensen, 2015) for photo-realistic rendering. We drive our head models with expression coefficients and head rotation angles estimated from our own recorded talking videos.

4.3 Evaluation Metrics

For evaluation metrics, we report peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and perceptual similarity (LPIPS) (Zhang et al., 2018) to measure the quality of generation results. For datasets, we collect 8 talking portrait videos from YouTube with an average length of around 3 min as our real video set (most are used in AD-NeRF (Guo et al., 2021) or HDTF (Zhang et al., 2021c)) and additionally render synthetic videos of 6 persons with an average length of around 2 min for quantitative comparison. More details are introduced in our supplementary materials. To measure audio-driven accuracy, we further use SyncNet (confidence) (Chung & Zisserman, 2017) to measure the audio-driven synchronization.

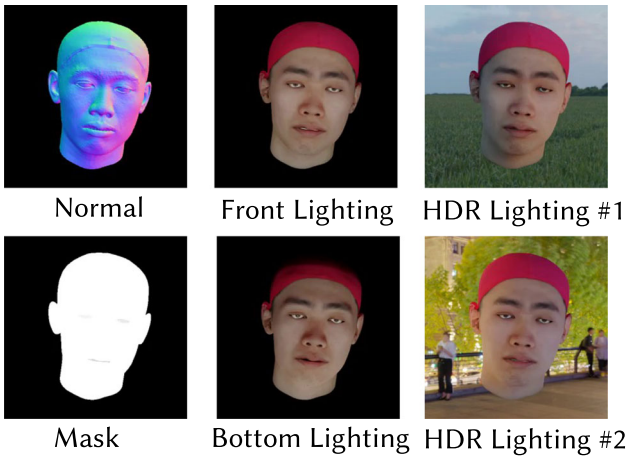

Fig. 4 Visualization of Synthetic Data. We render 6 sequences (2 min, 25 fps), for each person in 10 different lighting conditions with Cycles renderer (Hess, 2013) in Blender (Community, 2018)

Fig. 5 Qualitative Comparison of Real Video Driving. Our method successfully drives the motion of lips. Compared to AD-IMAvatar (Zheng et al., 2022), AD-NeRF (Guo et al., 2021), and DFRF (Shen et al., 2022), our generated lip motion are closer to the ground truth (zoom in for a better view)

4.4 Qualitative Comparison

Currently, there is no unified framework for relightable audio-driven talking portrait generation. Therefore, we first compare our method with audio animation methods and relighting methods separately, then trivially combine two existing frameworks as the baseline of relightable audio-driven talking portrait generation.

Comparison on Audio-Driven Talk In this work, we focus on personalized animation, which only uses one video for training. We choose two representative personalized audio animation methods, AD-NeRF (Guo et al., 2021), and DFRF (Shen et al., 2022). We also modify the FLAME-based method IMAvatar (Zheng et al., 2022) to gain a simple audio-driven version, AD-IMAvatar as an additional baseline. In Fig. 5, AD-IMAvatar only generates coarse talking portraits with blurry teeth areas. And AD-NeRF is prone to generate artifacts at the junction of the neck and head. Compared to the results of AD-NeRF and DFRF, the motion of our generated lips is closer to the ground truth. Notably, our framework succeeds to generate clear teeth areas.

Comparison on Relighting In this paper, we compare our relighting performance with five advanced methods. DPR (Zhou et al., 2019a), SMFR (Hou et al., 2021) and GCFR (Hou et al., 2022) are trained on publicly available data and release their models. Since nearly none of One-Light-at-A-Time (OLAT) based methods release their code or models. We requested the authors to inference their models on our provided inputs (SIPR-W (Wang et al., 2020) and TR (Pandey et al., 2021)), and take results for comparisons (Fig. 6).

As presented in Fig. 7, although both DPR and SMFR are able to reflect given lighting conditions on generated images when a sample directional light is given, their generated portraits lack the special texture of a real human face. This is mainly because they do not account for model specular, which is a significant and noticeable feature of the human face. Meanwhile, the recent method GCFR is easy to produce unnatural shadows. In contrast, ReliTalk renders realistic human portraits with reserved facial details.

Additionally, when complex lighting optimized from HDR images is used (as shown in Fig. 6), DPR and SMFR tend to produce unsatisfactory results, many of which are not relevant to the given lighting conditions. And SMFR even fails to reconstruct some faces. SIPR-W will generate some relighted results whose color is similar to the background but can not reflect the varied lighting on the face. Although TR succeeds to generate some vivid relighted results, it loses some facial details and also mildly hurt the original identity. However, our framework performs well on both types of lighting and successfully renders the specular texture of the human face. This enables our generated avatar to blend in seamlessly with various backgrounds, as long as HDR data of the background is available, by matching the shading and lighting of the avatar to that of the background.

4.5 Quantitative Comparison

Evaluation on Audio-Driven Talk As shown in Table 1, we compare our method with AD-IMAvatar, AD-NeRF, and

Fig. 6 Qualitative Comparisons of Real Video Relighting. We compare our methods against DPR (Zhou et al., 2019a), SMFR (Hou et al., 2021), SIPR-W (Wang et al., 2020) and TR (Pandey et al., 2021). ReliTalk renders human portraits with high-fidelity even with the complex lighting from HDR environment maps

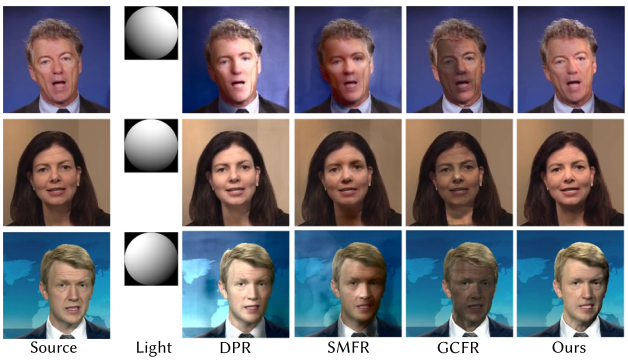

Fig. 7 Qualitative Comparisons Under Directional Lights. We compare our methods against three baseline methods DPR (Zhou et al., 2019a), SMFR (Hou et al., 2021) and GCFR (Hou et al., 2022) for portrait relighting under directional light

DFRF. Among those baselines, DFRF achieves comparable performance in PSNR, SSIM, and LPIPS, while its confidence in SyncNet is slightly lower than AD-NeRF. However, our method significantly outperforms all baselines in terms of all metrics.

Although our method uses some portrait area from existing frames, both AD-NeRF and DFRF use pose parameters from the existing sequence. In this way, DFRF only generates the remaining face area with the neck and collar given. Therefore for a fair comparison, we recalculate PSNR, SSIM, and LPIP purely in the driven area (120 × 120 resolution). As shown in the brackets of Table 1, our method still significantly outperforms all baselines in terms of all metrics.

Evaluation on Synthetic Relighting Dataset As shown in Table 2, we achieve the highest PSNR and SSIM on the synthetic dataset. And they are significantly higher than the other two, which indicates our generated images is closer to the ground truth. Meanwhile, the lowest LPIPS also illustrates that our results have the highest perceptual quality. It is notable that SMFR almost fails in our synthetic video dataset, which is perhaps caused by the distribution gap between our synthesized video data and real data. As a result, our ReliTalk outperforms DPR and SMFR both qualitatively and quantitatively.

According to the analysis of practicality, DPR and SMFR need a pre-collect face image dataset to train the network. At the inference stage, a new lighting or a new portrait that is out of training distribution will significantly hurt the performance. While our method may not be able to handle all persons within a single model, it is still practical because the training data, a short talking portrait video, is readily available and easy to obtain.

4.6 Ablation of Core Modules

Effects of Mesh-Aware Guidance Mesh-aware guidance is used to assist lip-syncing. We use a model pre-trained (Cudeiro et al., 2019b) to gain lip-related meshes as additional guidance. Phoneme-related features and lip-related meshes are separately encoded and then concatenated to generate pose and expression coefficients. As shown in Table 3, mesh-aware guidance improves prediction accuracy significantly. In addition, we find that the improvement is significant (PSNR increases from 29.5445 to 34.8186) when the training video is too short to cover enough phonemes (2450 training frames). The improvement is mild for the long video (6500

Table 1 Quantitative results of audio-driven real videos

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	SyncNet ↑
AD-IMAvatar (Zheng et al., 2022)	25.0625	0.8885	0.0538	2.5428
AD-NeRF (Guo et al., 2021)	25.6916 (29.8458)	0.9219 (0.9750)	0.1165 (0.0594)	3.8616
DFRF (Shen et al., 2022)	33.2088 (33.9563)	0.9665(0.9834)	0.1178 (0.0616)	3.7190
Ours	**37.6645 (37.9082)**	**0.9892 (0.9931)**	**0.0028 (0.0029)**	**5.5343**
Ground Truth	–	1.000	0.000	7.7218

Bold values indicate the best performance in the comparison
Our method significantly outperforms all baselines in terms of all metrics

Table 2 Quantitative results of synthetic video relighting

Methods	PSNR ↑	SSIM ↑	LPIPS ↓
DPR (Zhou et al., 2019a)	18.1899	0.9093	0.0668
SMFR (Hou et al., 2021)	15.9565	0.8003	0.3358
Ours	**22.8152**	**0.9435**	**0.0326**

Bold values indicate the best performance in the comparison
Our method achieves the highest PSNR and SSIM on the synthetic dataset

Table 3 Ablation results of Mesh-Aware Guidance

Methods	PSNR ↑	SSIM ↑
Audio Only	29.6147	0.9511
Mesh Only	33.8643	0.9790
Audio + Mesh (Ours)	**34.1099**	**0.9802**

Bold values indicate the best performance in the comparison
Our mesh-aware guidance improves prediction accuracy significantly

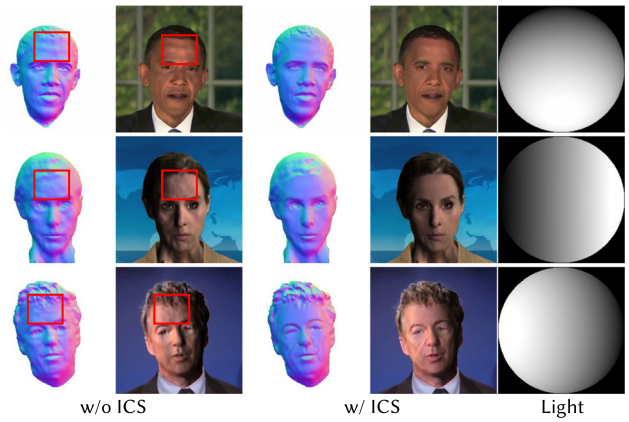

Fig. 8 Finer Normal under Identity-Consistent Supervision with Relighting. Without ICS, the normal map estimation may contain severe artifacts that cause unrealistic rendering given novel illuminations (zoom in for clearer results)

Table 4 Ablation Results of Identity-Consistent Supervision.

Methods	PSNR ↑	$\text{PSNR}_{\text{grad}}$ ↑	SSIM ↑
w/o ICS	21.7141	21.9828	0.9060
w ICS	21.5835	23.5441	0.9203

Without identity-consistent supervision, the estimated normal map will become noisy and contain more artifacts, indicated by a higher error in the gradient map

training frames). This implies that mesh-aware features offer the network approximate information about the movements of the lips, such as opening and closing, even when the input audio is significantly different from the audio used during training.

Effects of Identity-Consistent Supervision Identity-consistent supervision with relighting is employed to lessen the influence of lacking multi-view information. We visualize the effects in Fig. 8. Instead of a well-structured normal, the network prefers to generate an irregular one whose surface varies with the change of color on the face because it is an easier mapping for the coarse render. However, those irregular areas will be very significant when a different lighting is given (left of Fig. 8). After adding identity-consistent supervision, this weird face is hard to be recognized as the same person, urging the network to produce a well-structured normal which can gain reasonable relighting results under lighting with various directions (right of Fig. 8). To quantitatively evaluate the improvements of identity-consistent supervision, we calculate metrics in the image space of the normal map. Here we propose $\text{PSNR}_{\text{grad}}$, which calculates the 2D gradient in the image space, to jointly measure the normal quality. As shown in Table 4, although normal refined by ICS does not gain a higher PSNR, its $\text{PSNR}_{\text{grad}}$ is significantly higher, which means that it owns a better shape surface. Higher SSIM also proves the effectiveness of our method.

4.7 Ablation of Reflectance Decomposition

To get rid of leveraging knowledge from expensive capturing data (*i.e.* Light Stage (Debevec et al., 2000)), we decompose the human portrait into corresponding maps from monocular videos through some careful designs.

Initial Normal Different from previous audio-driven generation methods only generate the final portrait image, our audio-to-geometry also provides a good initial normal esti-

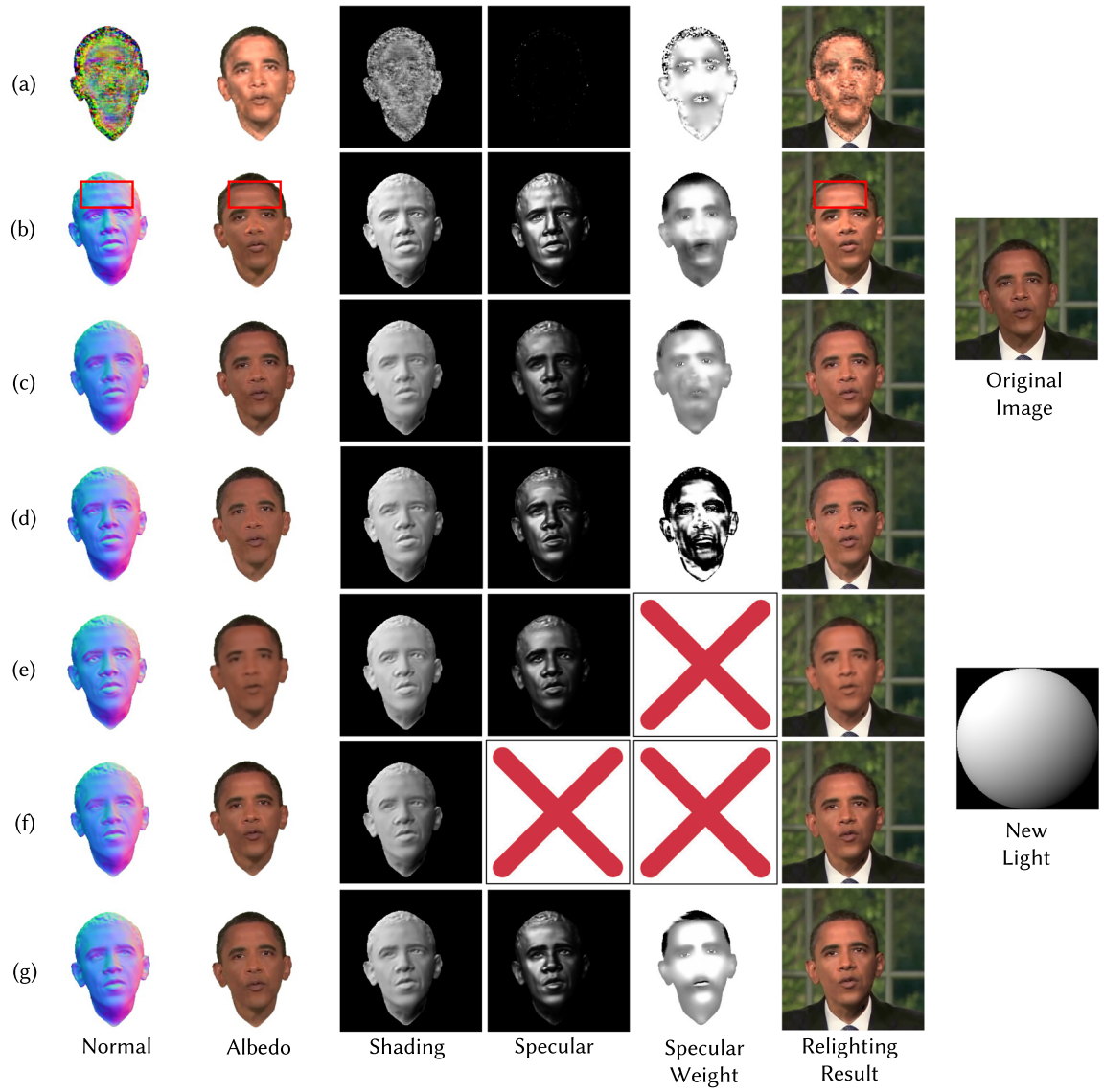

Fig. 9 Ablation of Reflectance Decomposition. **(a)** Without initial normal, **(b)** without normal residual, **(c)** without parsimony prior, **(d)** without smoothness constraints, **(e)** without specular weight, **(f)** without specular map, **(g)** our final method. Our final method gains the most vivid relighting results

mation. As shown in row (**a**) of Figs. 9, 10, 11, the reflectance decomposition can not converge properly without initial normal estimation because of lacking the constraints for the normal map.

Normal Residual Initial normal is not accurate because we do not have either the ground truth of the normal map or multi-view information of the portrait. Those irregular areas will be very significant when new lighting is given (row **(b)** of Fig. 9).

Parsimony Prior Parsimony means that the palette with which an albedo image was painted should be small. Without parsimony prior, albedo will contain more details while normal details are reduced, which is obviously reflected in the shading map of row **(c)** (Fig. 9). Therefore, the final relighting result is not such vivid.

Smoothness Constraints With smoothness constraints, some decomposition maps may overfit training data. As shown in row **(d)** of Fig. 9, the predicted specular weight is chaotic, reducing the vividness of the relighting result.

Specular Weight We use a network F_{cspec} to predict a specular weight C_{spec} for flexibly adjusting the final specular. As shown in row **(e)** of Fig. 9, the relighting result is blurred without this design.

Specular Map Prior works (He et al., 2016; Shu et al., 2017) for portrait relighting only employ simple diffuse lighting to model the human face. However, ignoring specular phe-

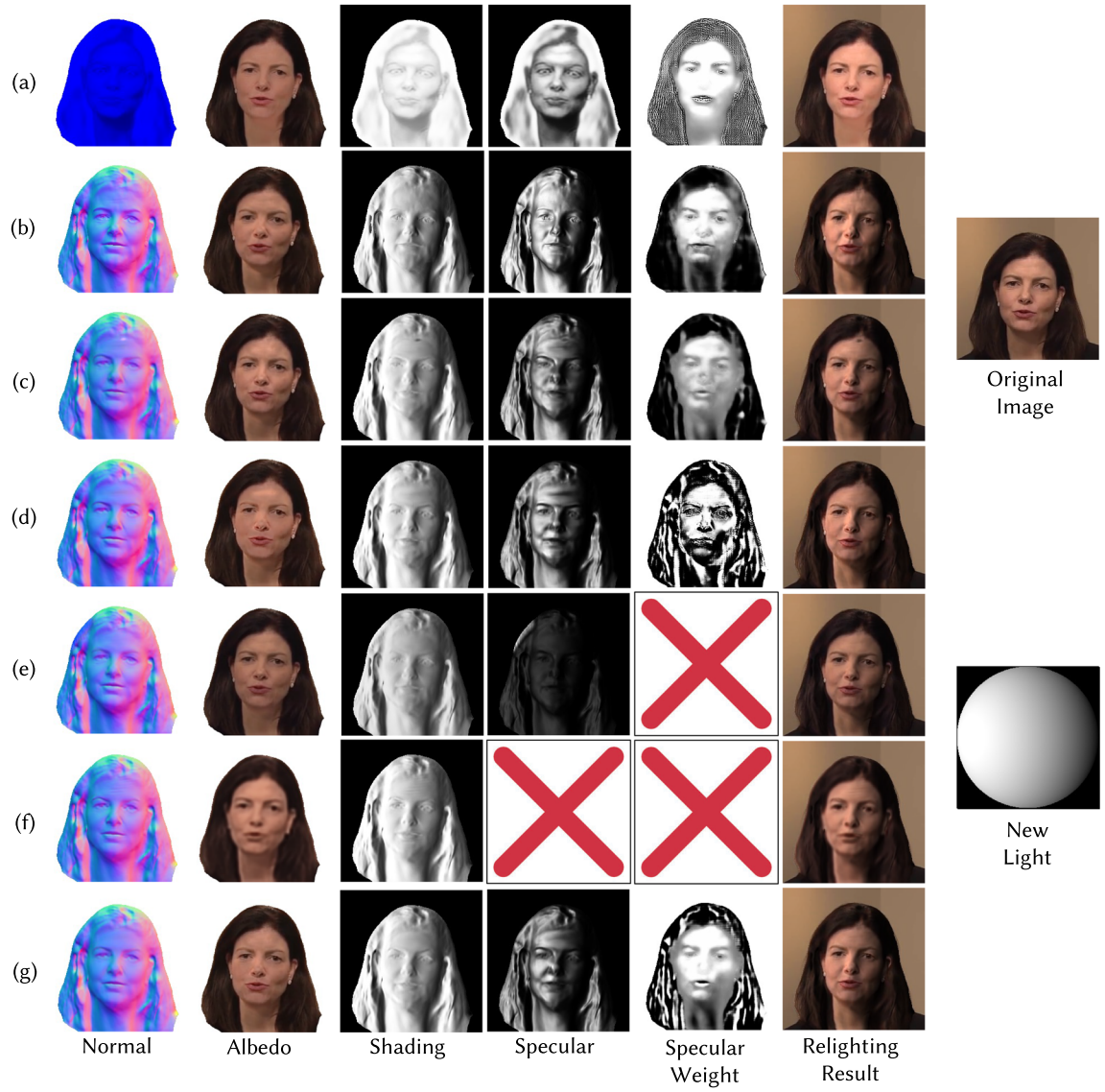

Fig. 10 Ablation of Reflectance Decomposition. **a** Without initial normal, **b** without normal residual, **c** without parsimony prior, **d** without smoothness constraints, **e** without specular weight, **f** without specular map, **g** our final method. Our final method gains the most vivid relighting results

nomenons that widely appear on human faces, the final rendering result is less photo-realistic (row **(f)** of Fig. 9).

4.8 Ablation of Training Frames

We also conduct a convergence ablation based on the number of training frames. As shown in Fig. 12, 750 training frames (clip of 30 s) are enough for basic relighting performance. But more training frames will bring better results. In this example, the unnatural division between hair and forehead gradually disappears when more training frames are used.

5 Conclusion

We propose **ReliTalk** a novel framework for relightable audio-driven talking portrait generation which only requires an easily accessible single monocular portrait video as input, while previous light-stage-based methods are not publicly available for data or code. Our method decomposes the human portrait for the well-disentangled geometry and reflectance, which is also expression- and pose-related. During the inference, we use audio from the user to drive the human portrait by controlling expression and pose coefficients, then render it with any desired illuminations, seamlessly harmonizing with the background.

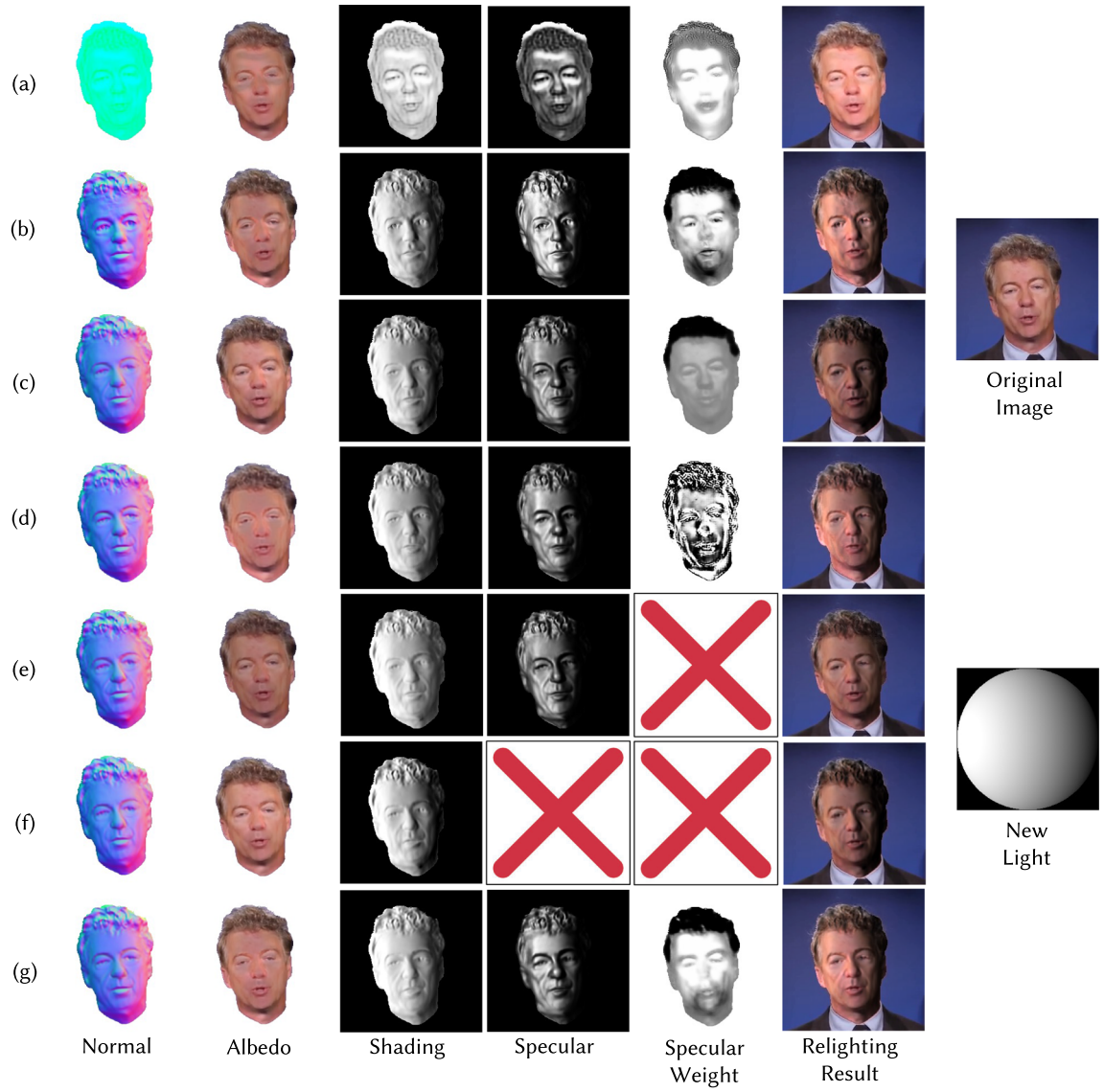

Fig. 11 Ablation of Reflectance Decomposition. **a** Without initial normal, **b** without normal residual, **c** without parsimony prior, **d** without smoothness constraints, **e** without specular weight, **f** without specular map, **g** our final method. Our final method gains the most vivid relighting results

However, there are still some limitations of our designed relighting model. (1) We only consider one-bounce direct environment light, and thus our method cannot handle furry appearances, such as beards and long hair. (2) Our method assumes the appearance of human faces does not change throughout the entire video. Therefore, actions like wearing glasses or putting on hats may change the appearance would cause inaccurate estimation of reflectance.

In the future, we want to design a more realistic physical model that can take into account various complex lighting conditions.

Societal Impacts Our code is released for better promotion. Therefore, users only need to input a talking video of the target person and then are able to freely generate a talking portrait with desired audio and background. Although it increases the risk of forged videos, our approach also provides a new type of forged samples for researchers to improve defense methods.

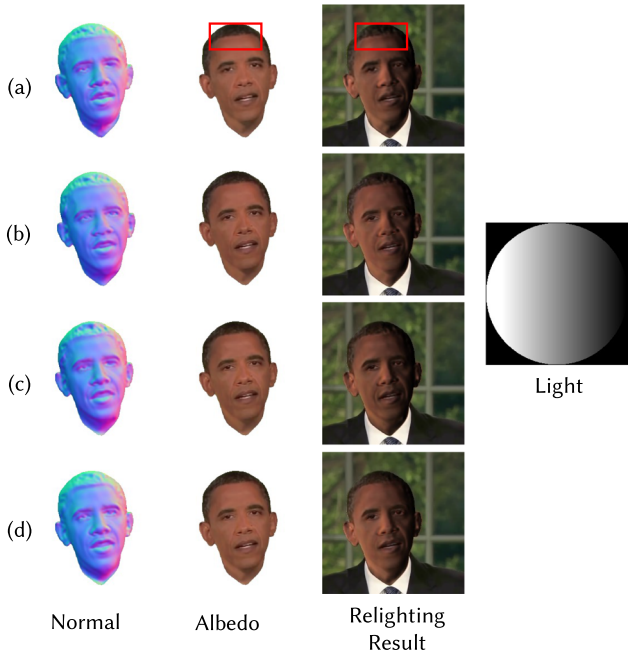

Fig. 12 Ablation of Training Frames. **a** 750 training frames, **b** 1500 training frames, **c** 3000 training frames, **d** 6000 training frames. 750 training frames (clip of 30 s) are enough for basic relighting performance. But more training frames will bring better results. In this example, the unnatural division between hair and forehead gradually disappears when more training frames are used

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02007-9.

Acknowledgements This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2022-01-035T), NTU NAP, MOE AcRF Tier 2 (MOET2EP20221-0012), and under the RIE2020 Industry Alignment Fund - Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

Data Availability Statement Video data and pre-trained models used in this paper are available online. We provide corresponding source links for reproduction purposes in the **ReliTalk** repository https://github.com/arthur-qiu/ReliTalk.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., & Chen, G., et al. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182.

Barron, J. T., & Malik, J. (2014). Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(8), 1670–1687.

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., & Srinivasan, P. P. (2021a). Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. arXiv:2103.13415 [cs].

Barron, J. T., Mildenhall, B., Verbin, D., Srinivasan, P. P., & Hedman, P. (2021b). Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. arXiv:2111.12077 [cs].

Basri, R., & Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(2), 218–233.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194.

Blinn, J. F. (1977). Models of light reflection for computer synthesized pictures. In *Proceedings of the 4th annual conference on computer graphics and interactive techniques*, pp. 192–198.

Caselles, P., Ramon, E., Garcia, J., Giro-i Nieto, X., Moreno-Noguer, F., & Triginer, G. (2023). Sira: Relightable avatars from a single image. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 775–784.

Chan, E.R., Monteiro, M., Kellnhofer, P., Wu, J., & Wetzstein, G. (2021). pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5799–5809.

Chan, E. R., Lin, C. Z., Chan, M. A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L. J., Tremblay, J., & Khamis, S., et al. (2022). Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16123–16133.

Chen, L., Maddox, R. K., Duan, Z., & Xu, C. (2019). Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7832–7841.

Chen, Z., & Liu, Z. (2022). Relighting4d: Neural relightable human from videos. In *European conference on computer vision*, Springer, pp. 606–623.

Christensen, P. H. (2015). An approximate reflectance profile for efficient subsurface scattering. In *ACM SIGGRAPH 2015 Talks*, pp. 1–1.

Chung, J. S., & Zisserman, A. (2017). Out of time: Automated lip sync in the wild. In *Computer Vision—ACCV 2016 Workshops: ACCV 2016 International Workshops*, Taipei, Taiwan, November 20–24, 2016, Revised Selected Papers, Part II 13, Springer, pp. 251–263.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. In *IEEE conference on computer vision and pattern recognition*.

Community, B. O. (2018). *Blender: A 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, http://www.blender.org.

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., Black, M. J. (2019a). Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10101–10111.

Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., & Black, M. J. (2019b). Capture, learning, and synthesis of 3d speaking styles. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10093–10103, https://doi.org/10.1109/CVPR.2019.01034.

Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., & Sagar, M. (2000). Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., USA, SIGGRAPH '00, pp. 145–156, https://doi.org/10.1145/344779.344855.

Fan, Y., Lin, Z., Saito, J., Wang, W., & Komura, T. (2022). Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG), 40*(4), 1–13.

Guo, Y., Chen, K., Liang, S., Liu, Y. J., Bao, H., & Zhang, J. (2021). Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5784–5794.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hess, R. (2013). *Blender foundations: The essential guide to learning blender 2.5*. Routledge.

Hou, A., Zhang, Z., Sarkis, M., Bi, N., Tong, Y., & Liu, X. (2021). Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14719–14728.

Hou, A., Sarkis, M., Bi, N., Tong, Y., & Liu, X. (2022). Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4217–4226.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C.C., Cao, X., Xu, F. (2021). Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14080–14089.

Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans Graph, 36*(4), https://doi.org/10.1145/3072959.3073658.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., & Theobalt, C. (2018). Deep video portraits. *ACM Transactions on Graphics (TOG), 37*(4), 1–14.

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Trans Graph, 36*(6), 194-1.

Liu, X., Xu, Y., Wu, Q., Zhou, H., Wu, W., & Zhou, B. (2022). Semantic-aware implicit neural audio-driven video portrait generation. arXiv preprint arXiv:2201.07786.

Liu, Y., Li, Y., You, S., & Lu, F. (2019). Unsupervised learning for intrinsic image decomposition from a single image. https://doi.org/10.48550/ARXIV.1911.09930.

Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. arXiv:2003.08934 [cs] version: 2.

Nestmeyer, T., Lalonde, J. F., Matthews, I. & Lehrmann, A. (2020). Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5124–5133.

Or-El, R., Luo, X., Shan, M., Shechtman, E., Park, J. J. & Kemelmacher-Shlizerman, I. (2022). Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13503–13513.

Pan, X., Dai, B., Liu, Z., Loy, CC. & Luo, P. (2020). *Do 2D gans know 3d shape? Unsupervised 3D shape reconstruction from 2d image gans*. arXiv preprint arXiv:2011.00844.

Pandey, R., Escolano, S. O., Legendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., & Fanello, S. (2021). Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG), 40*(4), 1–21.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British Machine vision conference*.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., & Jawahar, C. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492.

Ramamoorthi, R., & Hanrahan, P. (2001). On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A, 18*(10), 2448–2459.

Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., & Sheikh, Y. (2021). Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International conference on computer vision (ICCV)*, pp. 1153–1162, https://doi.org/10.1109/ICCV48922.2021.00121.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Shen, S., Li, W., Zhu, Z., Duan, Y., Zhou, J., & Lu, J. (2022). Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*.

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., & Samaras, D. (2017). Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5541–5550.

Song, L., Wu, W., Qian, C., He, R., & Loy, C. C. (2022). Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security, 17*, 585–598.

Srinivasan, P. P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., & Barron, J. T. (2021). Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*.

Sun, J., Wang, X., Zhang, Y., Li, X., Zhang, Q., Liu, Y., & Wang, J. (2022). Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7672–7682.

Sun, T., Barron, J. T., Tsai, Y. T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P. E., & Ramamoorthi, R. (2019). Single image portrait relighting. *ACM Transactions on Graphics (TOG), 38*(4), 79-1.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (ToG), 36*(4), 1–13.

Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., Liu, J., Zeng, G., & Wang J. (2022). Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. arXiv preprint arXiv:2211.12368

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., & Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG), 36*(4), 1–11.

Thies, J., Elgharib, M., Tewari, A., Theobalt, C., & Nießner, M. (2020). Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, Springer, pp. 716–731.

Wang, Y., Zhang, L., Liu, Z., Hua, G., Wen, Z., Zhang, Z., & Samaras, D. (2008). Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(11), 1968–1984.

Wang, Y., Holynski, A., Zhang, X., & Zhang, X. C. (2022). Sunstage: Portrait reconstruction and relighting using the sun as a light stage. arXiv preprint arXiv:2204.03648.

Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., & Xu, F. (2020). Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG), 39*(6), 1–13.

Wu, H., Jia, J., Wang, H., Dou, Y., Duan, C., & Deng, Q. (2021). Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1478–1486.

Xu, Y., Peng, S., Yang, C., Shen, Y., & Zhou, B. (2022). 3d-aware image synthesis via learning structural and textural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18430–18439.

Yang, H., Zhu, H., Wang, Y., Huang, M., Shen, Q., Yang, R., & Cao, X. (2020). Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Yao, S., Zhong, R., Yan, Y., Zhai, G., & Yang, X. (2022). *DFA-NeRF: Personalized talking head generation via disentangled face attributes neural rendering*. arXiv:2201.00791 [cs].

Yeh, Y. Y., Nagano, K., Khamis, S., Kautz, J., Liu, M. Y., Wang, T. C. (2022). Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. arXiv preprint arXiv:2209.10510.

Yi, R., Ye, Z., Zhang, J., Bao, H., & Liu, Y. J. (2020). Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363.

Zhang, L., Zhang, Q., Wu, M., Yu, J., & Xu, L. (2021a). Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 802–812.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*.

Zhang, X., Srinivasan, P. P., Deng, B., Debevec, P., Freeman, W. T., & Barron, J. T. (2021). Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG), 40*(6), 1–18.

Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021c). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3661–3670.

Zhao, X., Ma, F., Güera, D., Ren, Z., Schwing, A. G., & Colburn, A. (2022). Generative multiplane images: Making a 2d gan 3d-aware. In *European conference on computer vision*, Springer, pp. 18–35.

Zheng, Y., Abrevaya, V. F., Bühler, M. C., Chen, X., Black, M. J., & Hilliges, O. (2022). Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13545–13555.

Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D. W. (2019a). Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7194–7202.

Zhou, H., Liu, Y., Liu, Z., Luo, P., & Wang, X. (2019). Talking face generation by adversarially disentangled audio-visual representation. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 9299–9306.

Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., & Liu, Z. (2021). Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4176–4186.

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. (2020). Makelttalk: Speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG), 39*(6), 1–15.

Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Shen, Q., Yang, R., & Cao, X. (2021). Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. arXiv preprint arXiv:2111.01082.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.