VRMM: A Volumetric Relightable Morphable Head Model

Haotian Yang
Kuaishou Technology
China
yanght321@gmail.com

Mingwu Zheng
Kuaishou Technology
China
zhengmingwu@kuaishou.com

Chongyang Ma
Kuaishou Technology
China
chongyangm@gmail.com

Yu-Kun Lai
Cardiff University
United Kingdom
laiy4@cardiff.ac.uk

Pengfei Wan
Kuaishou Technology
China
wanpengfei@kuaishou.com

Haibin Huang*
Kuaishou Technology
China
jackiehuanghaibin@gmail.com

Figure 1: We present VRMM, a novel volumetric head prior with fully disentangled low-dimensional parametric space for identity, expression, and illumination. Trained on dynamic expressions of hundreds of people captured in a LightStage with controllable illumination, our VRMM enables high-quality animatable and relightable avatar reconstruction from few-shot observations.

ABSTRACT

In this paper, we introduce the Volumetric Relightable Morphable Model (VRMM), a novel volumetric and parametric facial prior for 3D face modeling. While recent volumetric prior models offer improvements over traditional methods like 3D Morphable Models (3DMMs), they face challenges in model learning and personalized reconstructions. Our VRMM overcomes these by employing a novel training framework that efficiently disentangles and encodes latent spaces of identity, expression, and lighting into low-dimensional representations. This framework, designed with self-supervised learning, significantly reduces the constraints for training data, making it more feasible in practice. The learned VRMM offers relighting capabilities and encompasses a comprehensive range of expressions. We demonstrate the versatility and effectiveness of VRMM through various applications like avatar generation, facial reconstruction, and animation. Additionally, we address the common issue of overfitting in generative volumetric models with a novel prior-preserving personalization framework based on VRMM. Such an approach enables high-quality 3D face reconstruction from even a single portrait input. Our experiments showcase the potential of VRMM to significantly enhance the field of 3D face modeling.

*Corresponding author.

SIGGRAPH Conference Papers '24, July 27-August 1, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07…$15.00
https://doi.org/10.1145/3641519.3657406

CCS CONCEPTS

• **Computing methodologies** → **Volumetric models**; **Motion capture**; **Reflectance modeling**.

KEYWORDS

3D avatar creation; Facial animation; Neural rendering; View synthesis; Relighting.

ACM Reference Format:

Haotian Yang, Mingwu Zheng, Chongyang Ma, Yu-Kun Lai, Pengfei Wan, and Haibin Huang. 2024. VRMM: A Volumetric Relightable Morphable Head Model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27-August 1, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657406

1 INTRODUCTION

In this study, we explore 3D face modeling from a representation learning perspective. The pursuit of realistic 3D facial representations is pivotal in both academic research and practical applications, encompassing digital entertainment, telepresence, and biometric identification. As a critical task in computer graphics and computer vision, this area has garnered sustained attention. Early works like 3D Morphable Models (3DMMs) [Blanz and Vetter 1999; Booth et al. 2016; Cao et al. 2013; Jiang et al. 2019; Li et al. 2017; Paysan et al. 2009; Vlasic et al. 2005; Yang et al. 2020] have set a benchmark due to their ability to encode faces in a parametric form. Their parametric nature allows for the manipulation of face identity, expressions and other attributes with relative ease, making 3DMMs a powerful tool in the field. However, they often struggle to achieve a truly lifelike fidelity, particularly in the subtleties of complex head components such as hair and the interior of the mouth. To overcome these limitations, research has shifted towards volumetric models [Bühler et al. 2023; Cao et al. 2022; Hong et al. 2022; Wang et al. 2022; Zhuang et al. 2022], which offer a more comprehensive representation of facial structures and promise enhanced realism.

Despite the advancements offered by volumetric models, there remains a significant gap in their capability. Our main observations are two-fold: first of all, existing data-driven volumetric models are not able to adequately model dynamic facial expressions or to simulate the effects of variable lighting conditions on the face. Existing animatable volumetric prior models either adopt supervised learning with decoupled input data of discrete identities and expressions [Hong et al. 2022; Zhuang et al. 2022], or require expensive and brittle preprocessing steps [Cao et al. 2022], which is tedious and not practical to scale up. These drawbacks also limit the further query of a continuous relightable model. Subsequently, there will be common overfitting issues when using generative models for downstream reconstruction tasks [Abdal et al. 2019; Tewari et al. 2020; Tov et al. 2021; Zhu et al. 2020], which is also observed in volumetric prior personalization [Bühler et al. 2023]. This drawback is that due to the limited training data, the inversion of input facial data into the volumetric modeling space is often problematic and leading to reconstruction below satisfactory, which also deteriorates the editability, i.e., relightable and animatable properties learned in the prior model.

To bridge this gap, we propose VRMM, a novel volumetric and parametric 3D face prior. VRMM is built upon volumetric primitives linked to the UV space of a base mesh [Lombardi et al. 2021], and adapt a physically-inspired appearance decoder [Yang et al. 2023] for relighting . It uses multi-identity mapping and an expression encoder to handle various expressions and lighting conditions. By explicitly encoding identity, expression and lighting into low-dimensional representations, VRMM learns to disentangle the associated latent spaces and can be trained in a self-supervised manner. Our design of VRMM effectively reduces the constraints of training data required by previous volumetric models and enables training with more flexible data collections. In practice, the training of our VRMM is based on a dataset comprising high-quality multi-view image sequences of fewer than 300 individuals. These individuals were captured exhibiting dynamic expressions within a LightStage under controllable lighting conditions.

We further adapt our VRMM model for various fitting and reconstruction tasks. As discussed above, fine-tuning is necessary for volumetric prior models to overcome the inversion issue [Bühler et al. 2023; Cao et al. 2022; Wang et al. 2022] at the cost of overfitting. Hence we propose a prior-preserving framework for model fitting. Specifically, we use an identity related regularization term to balance the VRMM's generation capacity and fitting accuracy. Our specifically designed framework allows high-quality avatars that are animatable and relightable to be created from few-shot captures, significantly outperforming existing baselines.

In summary, our contributions are:

- We present VRMM, the first 3D volumetric facial prior that is both continuously relightable and encompasses full range of expressions to the best of our knowledge.
- We propose a novel training framework to learn the disentangled parametric space of expression, identity and lighting for VRMM from dynamic multi-view image sequences captured under controllable light conditions.
- We propose a novel personalization method that is elaborately designed to keep the animatable and relightable properties provided by the prior, which enables high-fidelity avatar reconstruction from several or even one image.
- Extensive experiments demonstrate that VRMM can be used in various applications and outperforms previous methods.

2 RELATED WORK

In this section, we review closely related parametric 3D face models based on both traditional mesh-based representation and volumetric representation. We also discuss related methods of neural avatar reconstruction.

Parametric head models. Accurate modeling of 3D head geometry and appearance remains a significant challenge in both computer graphics and computer vision. Among various methodologies, 3D Morphable Models (3DMMs) emerged as a pioneering and systematic approach. Originally introduced by Blanz and Vetter [1999] and later evolved into multi-linear models [Vlasic et al. 2005], 3DMMs have been fundamental in modeling facial meshes and textures. They serve as a universal facial prior in diverse applications like face reconstruction and tracking from single-view images [Dou et al. 2017b; Thies et al. 2016; Zhu et al. 2017]. However, 3DMMs are limited by their linear and mesh-based nature, restricting the scope of shape and appearance modeling.

Recent advances in deep learning have sought to address these limitations, employing nonlinear methods for more refined facial modeling [Bagautdinov et al. 2018; Ploumpis et al. 2020; Tran and Liu 2018, 2019; Zhang et al. 2022; Zheng et al. 2022]. Despite these improvements, surface-based representations, including both meshes and implicit functions, still struggle to capture the complete complexity of human head components such as teeth, facial hair, eyes, and so on.

A significant shift in this domain is evident with the advent of neural volumetric rendering, particularly NeRF (Neural Radiance Field)-based techniques [Kerbl et al. 2023; Mildenhall et al. 2020; Müller et al. 2022]. These methods holistically represent the 3D objects, achieving both photorealism and 3D consistency from multi-view images. Early attempts in volumetric head modeling

Table 1: Comparison of different methods for volumetric head models across four aspects. VRMM is the only model that enables a physically relightable and animatable model with real-time rendering capabilities. Single-stage reconstruction indicates that the morphable model does not require reconstructed meshes or other 3DMMs when fitting new data. (* Zhuang *et al.* [2022] and Hong *et al.* [2022] only support limited expressions. Hong *et al.* [2022] is limited to relighting under preset lighting conditions.)

Method	Animatable	Relightable	Single-stage reconstruction	Real-time rendering
Zhuang *et al.* [2022]	✓*	✗	✓	✗
Cao *et al.* [2022]	✓	✗	✗	✓
Hong *et al.* [2022]	✓*	✓*	✗	✓
Wang *et al.* [2022]	✗	✗	✓	✗
Buhler *et al.* [2023]	✗	✗	✓	✗
VRMM (Ours)	✓	✓	✓	✓

primarily focused on the geometry and appearance of a single scene or identity [Lombardi et al. 2019, 2021; Ma et al. 2021; Yang et al. 2023]. More recent work extends to multi-identity scenarios through generative modeling [Bühler et al. 2023; Cao et al. 2022; Hong et al. 2022; Wang et al. 2022; Zhuang et al. 2022]. For instance, MoRF [Wang et al. 2022] employs an auto-decoder framework [Park et al. 2019] to learn a conditional NeRF in a polarization-based studio setup, separating the diffuse and specular shading components. However, its applications are confined to studio environments and lack generalization to in-the-wild scenes. Preface [Bühler et al. 2023] expands on this, allowing high-resolution rendering in more casual settings. Nonetheless, both MoRF and Preface are limited to static head models, missing out on dynamic and relightable features which are crucial for real-world applications. MoFaNeRF [Zhuang et al. 2022] and Cao *et al.* [2022] integrate dynamic expressions, enabling the creation of animatable avatars. HeadNeRF [Hong et al. 2022] also introduces basic relighting capabilities but is constrained to certain lighting conditions. Currently, no previous method achieves full creation of expressive, dynamically relightable heads under various physically accurate lighting conditions. For a more comprehensive comparison, please refer to Tab. 1. To our knowledge, VRMM is the first model to achieve this, presenting a significant advancement in the field of volumetric and relightable morphable head models.

Avatar reconstruction. The field of 3D face reconstruction and performance capture has seen extensive research efforts over decades, leading to the development of sophisticated 3D scanning systems. These systems, focusing on static geometry reconstruction [Beeler et al. 2010; Ghosh et al. 2011] and dynamic performance capture [Beeler et al. 2011; Bradley et al. 2010; Collet et al. 2015; Dou et al. 2017a; Guo et al. 2019; Huang et al. 2011], predominantly employ multi-view stereo (MVS) and structured light techniques for point cloud acquisition. Subsequent steps involve estimating deforming geometry to maintain temporal mesh consistency. However, such tracking process often requires labor-intensive MVS reconstruction for numerous frames and complex optical-flow optimization, while real-time face tracking algorithms still fall short in accuracy.

Another vital component for creating realistic and relightable avatars is the estimation of light interaction with the subject, particularly the reflectance properties. Traditional methods typically model this interaction using bidirectional reflectance distribution functions (BRDF) [Schlick 1994], determined by observing appearance changes under various lighting conditions. Active lighting approaches, such as those using LightStage setups [Debevec et al. 2000], involve data collection with complex arrangements like one-light-at-a-time (OLAT) capture or polarized and color gradient illuminations [Fyffe and Debevec 2015; Ghosh et al. 2011; Guo et al. 2019; Ma et al. 2007; Zhang et al. 2022]. Conversely, passive capture methods [Li et al. 2022; Riviere et al. 2020; Zheng et al. 2023a] significantly reduce the need for elaborate setups.

Despite these advancements, these methods still require high hardware costs and considerable effort for data acquisition. The advent of deep learning and large-scale datasets has enabled the estimation of geometry and reflectance properties from single-view input [Caselles et al. 2023; Lattas et al. 2020, 2021; Li et al. 2020b,a; Papantoniou et al. 2023; Yamaguchi et al. 2018]. Nevertheless, these methods typically focus only on skin regions, as the complexity of hair and eye structures makes single-view inverse rendering difficult. Moreover, most reconstructed results are static and not animatable. Recent endeavors have focused on creating dynamic avatars from monocular videos [Chen and Liu 2022; Gao et al. 2022; Zheng et al. 2023b; Zielonka et al. 2023] or RGB-D inputs [Cao et al. 2022]. However, these methods typically fall short in providing relightable attributes, or their quality is constrained by the training video's limitations, due to the lack of a powerful prior model. The challenge remains to photorealistically model complete human heads, considering all the highly complex and diverse compositions of material, geometry, and expression.

To our best knowledge, our VRMM is the first capable of reconstructing high-quality animatable and relightable volumetric avatars from few-shot captures, representing a significant leap in the field of 3D avatar reconstruction.

3 METHOD

In this section, we first introduce the training framework of our volumetric relightable morphable model using high-quality data captured in the studio, as shown in Figure 2. Then we present a specially designed pipeline to fit the learned VRMM model to in the studio or in-the-wild captures for consumer-grade authentic 3D avatar reconstruction.

3.1 Preliminaries

Our VRMM is built upon the representation of Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021], which represents the scene as volumetric primitives attached to the UV space of a base mesh. We adopt the physically-inspired appearance decoder [Yang et al. 2023] to support relighting, where the lighting condition l is represented as the incoming light field of N_l densely sampled directions on the sphere. Specifically, given the expression code z_e, the view direction $\mathbf{d}$, and the lighting condition l, a series of decoders $\mathcal{D}_{MVP}$ predict the position and color of N_{prim} volumetric primitives for rendering:

$$\{\mathbf{v}, R_p, t_p, s_p, V_\alpha, V_{rgb}\} = \mathcal{D}_{MVP}(z_e, \mathbf{d}, l), \tag{1}$$

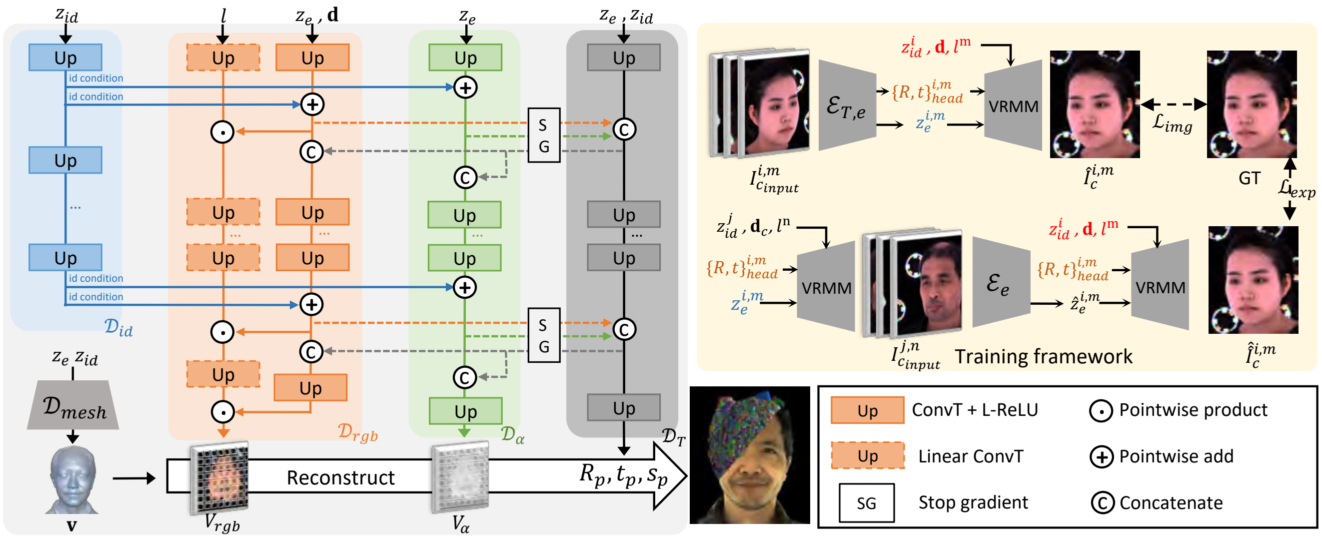

Figure 2: The VRMM pipeline. Network architecture (left): VRMM accepts inputs of identity code z_{id}, expression code z_e, view direction d, and environmental light l. The output, comprising a base mesh and volumetric primitives, is generated by respective decoders and rendered into an image in real-time. Notably, the transformation decoder $\mathcal{D}_T$, opacity decoder $\mathcal{D}_\alpha$, and the non-linear branch of the relightable appearance decoder $\mathcal{D}_{rgb}$ are interconnected through a detach-concatenation process between blocks, a key factor we found for achieving stable results. Training Framework (right): Our framework jointly trains the expression encoder $\mathcal{E}_e$, transformation encoder $\mathcal{E}_T$, per-person identity codes z_{id}, and the decoders in VRMM. Additionally, we incorporate a novel expression consistency loss $\mathcal{L}_{exp}$ to enhance the semantic alignment of expression codes.

where $\mathbf{v}$ is the position of the vertices of the base mesh, $\{R_p, t_p, s_p\}$ are the rotation, translation and non-uniform scaling of N_{prim} primitives relative to the base mesh, V_α and V_{rgb} are the opacity and color of each voxel in the primitives, respectively. The color of the rendered image I_{rgb} at pixel p can be obtained by integrating the radiance of the voxels along the direction $\mathbf{d}_p$ of the ray starting from the position $\mathbf{o}_p$ of pixel in 3D space:

$$I_{rgb}(p) = \int_{t_{\min}}^{t_{\max}} V_{rgb}(\mathbf{o}_p + t\mathbf{d}_p) \frac{dT(p,t)}{dt}, \tag{2}$$

$$T(p,t) = \min\left(1, \int_{t_{\min}}^{t} V_\alpha(\mathbf{o}_p + t\mathbf{d}_p)\right), \tag{3}$$

where $t_{\min}$ and $t_{\max}$ are the predetermined near and far range of the camera plane.

The volumetric primitives in MVP have consistent structure in the UV space of the base mesh, which improves the quality when extending to multiple identities as it is easier for the network to learn the shared features across identities compared to other alternatives based on the NeRF representation [Bühler et al. 2023; Hong et al. 2022; Wang et al. 2022; Zhuang et al. 2022].

3.2 VRMM

We start by extending the existing person-specific relightable MVP with a low-dimensional identity code z_{id} for our generative volumetric morphable head model. Then we discuss our modification to the network architecture and training process for disentanglement and better reconstruction quality.

Multi-identity model. Formally, our VRMM maps the identity code z_{id}, the expression code z_e, the view direction $\mathbf{d}$, and the lighting condition l to the base mesh and corresponding volumetric primitives:

$$\{\mathbf{v}, R_p, t_p, s_p, V_\alpha, V_{rgb}\} = \mathrm{VRMM}(z_{id}, z_e, \mathbf{d}, l). \tag{4}$$

Specifically, VRMM is composed of five decoders:

$$\mathrm{VRMM} := \{\mathcal{D}_{mesh}, \mathcal{D}_{id}, \mathcal{D}_T, \mathcal{D}_\alpha, \mathcal{D}_{rgb}\}. \tag{5}$$

The mesh decoder $\mathcal{D}_{mesh}$ is a multi-layer perceptron that predicts the vertex positions $\mathbf{v}$ of the base mesh given the identity code z_{id} and expression code z_e. The convolutional identity decoder $\mathcal{D}_{id}$ maps the low-dimensional identity code z_{id} to hierarchical feature maps $\mathcal{F}_{\text{alpha}}$ and $\mathcal{F}_{\text{appe}}$ that are injected to the opacity decoder $\mathcal{D}_\alpha$ and appearance decoder $\mathcal{D}_{rgb}$, respectively. The transformation decoder $\mathcal{D}_T$ maps the identity code z_{id} and expression code z_e to the rotation R_p, translation t_p, and scale s_p of the primitives for identity-related expression decoding. The opacity decoder $\mathcal{D}_\alpha$ predicts the voxel opacity V_α of the primitives conditioned on the expression code z_e and the feature maps $\mathcal{F}_{\text{alpha}}$ from the identity encoder. The relightable appearance decoder $\mathcal{D}_{rgb}$ adopts an architecture similar to [Yang et al. 2023] that takes view direction $\mathbf{d}$, lighting condition l, and expression code z_e as input to predict the voxel color V_{rgb} with a linear branch for lighting related decoding and $\mathcal{F}_{\text{appe}}$ hierarchically injected to the non-linear branch.

Considering the fact that the transformation and appearance of the volumetric primitives are closely related, we further concatenate the feature maps from the intermediate layers of the convolutional transformation decoder $\mathcal{D}_T$ to corresponding layers of the opacity decoder $\mathcal{D}_\alpha$ and the non-linear branch of the relightable appearance decoder $\mathcal{D}_{rgb}$, and vice-versa. However, direct concatenation leads to unstable convergence during training, which we believe is due to the scale discrepancy of gradient in different branches. So we stop the gradient back-propagation through the concatenated feature maps. We empirically find that the detach-concatenate operation significantly alleviate the jittering of the volumetric primitives when the numbers of identities and training frames increase.

Different from [Cao et al. 2022] where the meshes and textures are directly fed into the prior model for identity encoding, VRMM learns to generate novel relightable identities from the low-dimensional identity code, enabling VRMM to be used for avatar reconstruction by directly fitting to images without depending on traditional mesh-based parametric face models.

Disentangled latent space training. Different from previous volumetric head priors [Hong et al. 2022; Zhuang et al. 2022] that are trained on limited predefined expressions, our VRMM is designed to encompass a full range of dynamic expressions for animatable avatar reconstruction. Besides, the relightable appearance means capturing under varying lighting conditions, which leads to a significant challenge, as it is not feasible to capture many people performing a large number of identical expressions under different lighting conditions even in the studio.

Inspired by recent cross-identity neural retargeting methods [Xu et al. 2023; Zhang et al. 2022], we propose a novel framework for our VRMM to learn shared expression latent space across identities without explicit supervision. We also adopt the tracking-free training pipeline [Yang et al. 2023] that jointly learns topology consistent tracking end-to-end from dynamic image sequences, which avoids the brittle preprocessing step of surface-tracking (especially with changing lightings) and enables scalable training for a large number of different identities.

Specifically, given the training dataset with image $I_c^{i,m}$ of dynamic performance of subject i captured by camera c at frame m under known illumination l^m, we jointly train an expression encoder $\mathcal{E}_e$ that predicts the mean and variance of a multi-variant Gaussian distribution for expression code as well as a transformation encoder $\mathcal{E}_T$ that predicts the rigid rotation $R^{i,m}$ and translation $t^{i,m}$ of the head in that frame from a subset of camera views $I_{c_{input}}^{i,m}$. The expression code z_e is then sampled from the Gaussian distribution. The encoding process can be represented as:

$$z_e^{i,m} = \mathcal{E}_e(I_{c_{input}}^{i,m}), \tag{6}$$

$$R^{i,m}, t^{i,m} = \mathcal{E}_T(I_{c_{input}}^{i,m}). \tag{7}$$

As for the identity code, instead of using the full auto-encoder framework, we adopt the decoder-only method [Bühler et al. 2023; Wang et al. 2022] where the identity code z_{id}^i for each subject is initialized with Gaussian noise and jointly optimized with the networks during training. Note that the auto-decoder framework learns a generative model though there is not an explicit sampling process during training as discussed in [Park et al. 2019]. Then the synthesized image $\hat{I}_c^{i,m}$ of a camera c with view direction $\mathbf{d}_c$ and camera parameters ϕ_c is given by:

$$\hat{I}_c^{i,m} = \Pi(\text{VRMM}(z_{id}^i, z_e^{i,m}, \mathbf{d}_c, l^m), R^{i,m}, t^{i,m}, \phi_c), \tag{8}$$

where Π represents the aforementioned differentiable ray-marching process for rendering.

The loss $\mathcal{L}_{total}$ of the training objective function consists of three parts:

$$\mathcal{L}_{total} = \mathcal{L}_{img} + \mathcal{L}_{reg} + \mathcal{L}_{exp}, \tag{9}$$

where $\mathcal{L}_{img} = \mathcal{L}_1 + \lambda_{VGG}\mathcal{L}_{VGG}$ comparing the reconstruction $\hat{I}_c^{i,m}$ and the observed image $I_c^{i,m}$ consists of the L_1 loss term $\mathcal{L}_1$ and the perceptual loss term $\mathcal{L}_{VGG}$ with weight λ_{VGG}. The regularization loss is given by:

$$\mathcal{L}_{reg} = \lambda_{KLD}\mathcal{L}_{KLD} + \lambda_{vol}\mathcal{L}_{vol} + \lambda_{scale}\mathcal{L}_{scale} + \lambda_{id}||z_{id}||_2^2, \tag{10}$$

where $\mathcal{L}_{KLD}$ is the KL-divergence loss between the distribution of the expression code z_e and multi-variant Gaussian prior. $\mathcal{L}_{vol}$ is the volume minimization prior proposed in MVP [Lombardi et al. 2021]. The scale regularization term $\mathcal{L}_{scale}$ penalizes the k sides with the shortest length in the N_{prim} cubic primitives to prevent the volumetric primitives from squeezing without affecting normal primitives:

$$\mathcal{L}_{scale} = \sum_{n \in \mathcal{S}} \log(1/s_p^n), \tag{11}$$

where $\mathcal{S}$ is the set of indices of the k shortest sides among all the sides s_p, and s_p^n is the predicted scale of the n-th side. The $L2$ regularization term on identity code z_{id} is derived by assuming the prior distribution of z_{id} to be multi-variant Gaussian distribution. λ_{KLD}, λ_{vol}, λ_{scale}, and λ_{id} are balancing weights.

We find that the KL-divergence regularization $\mathcal{L}_{KLD}$ on the expression code plays an important role for our VRMM to learn a disentangled representation for identity and expression. By applying a much larger regularization weight λ_{KLD} on expression code compared to λ_{id}, the information in z_e is limited so that the expression encoder $\mathcal{E}_e$ learns to extract only expression-related information while lighting and identity are injected from other branches. We also experiment with the adversarial training for disentanglement [Schwartz et al. 2020; Zhang et al. 2022] but observe no improvement.

The system tends to learn a shared expression space even without a specific constraint as the decoders are shared across identities. To further align the semantic meaning for the expression code z_e for different identities, we propose to incorporate a novel expression consistency loss $\mathcal{L}_{exp}$. Specifically, given the expression code $z_e^{i,m}$ predicted by expression encoder $\mathcal{E}_e$ from $I_{c_{input}}^{i,m}$, we randomly choose an identity code z_{id}^j from another identity j and a different lighting condition l^n. Then we render images corresponding to $z_e^{i,m}$, z_{id}^j, and l^n:

$$\hat{I}_{c_{input}}^{j,n} = \Pi(\text{VRMM}(z_{id}^j, z_e^{i,m}, \mathbf{d}_{c_{input}}, l^n), R^{i,m}, t^{i,m}, \phi_{c_{input}}). \tag{12}$$

$\hat{I}_{c_{input}}^{j,n}$ should have the same expression as in $I_{c_{input}}^{i,m}$ but different identity and lighting condition. We feed $\hat{I}_{c_{input}}^{j,n}$ to the expression encoder $\mathcal{E}_e$ to extract the corresponding expression code $\hat{z_e}^{i,m}$. We also render the image $\hat{\hat{I}}_c^{i,m}$ with $\hat{z_e}^{i,m}$, the original identity z_{id}^i, and lighting l^m, as directly measuring the distance in the parameter space is shown to be inefficient [Tewari et al. 2020, 2017]. The expression consistency loss $\mathcal{L}_{exp}$ is then given by:

$$\mathcal{L}_{exp} = \lambda_{imgexp}\mathcal{L}_{imgexp} + \lambda_{parexp}||\hat{z_e}^{i,m} - z_e^{i,m}||_2^2, \tag{13}$$

Figure 3: Our model allows real-time global illumination. The lighting condition is represented as latitude-longitude environment maps, which is shown on the top.

where $\mathcal{L}_{imgexp}$ measures the image space difference similar to $\mathcal{L}_{img}$, λ_{imgexp} and λ_{parexp} are the weights of different terms.

3.3 Model Fitting

Once trained, VRMM can be used for avatar reconstruction with the analysis-by-synthesis scheme as traditional 3DMMs. However, volumetric avatars include complex components that cannot be expressed by the constrained parametric space of the prior model, as it was trained with only hundreds of identities. For faithful reconstruction, we further finetune VRMM for personalized avatar generation similar to [Bühler et al. 2023; Cao et al. 2022; Wang et al. 2022]. We find finetuning can increase the reconstruction quality but make manipulation deteriorate. Similar distortion-editability trade-off is also demonstrated in the GAN inversion field [Tov et al. 2021; Zhu et al. 2020]. Our fitting pipeline is specially designed to preserve the data-driven prior that can create a relightable and animatable avatar from even a single image. We illustrate our pipeline for fitting a single in-the-wild image for clarity, which can be easily extended to image sets.

Inversion. Given an image, we first detect 2D facial landmarks and optimize the rough rigid transformation and camera projection parameters with respect to the corresponding predefined 3D landmarks on the mean base mesh of VRMM for initialization. Then we jointly optimize the input parameters of VRMM and camera projection by inverse rendering. Particularly, instead of directly solving for the identity code z_{id}, we find a set of weights $\mathbf{w}$ that linearly blends the existing identity codes in the training set to constrain the domain. The objective function is formulated as:

$$\arg\min_{\mathbf{w},z_e,l,R,t,\phi} \mathcal{L}_{img} + \lambda_{exp}||z_e||_2^2, \tag{14}$$

where $\mathcal{L}_{img}$ is identical to the data term in Equation 9 and λ_{exp} is the weight of regularization term. The lighting l is restrict to be non-negative during optimization. The view direction $\mathbf{d}$ can be computed as inversion of the rotation R.

Figure 4: Qualitative comparison results on novel view synthesis. Our method produces more faithful results compared to existing parametric head models MoFaNeRF [Zhuang et al. 2022] and HeadNeRF [Hong et al. 2022].

Fine-tuning. After inversion we obtain an avatar that is animatable and relightable but may not authentically replicate the image. Inspired by Pivotal Tuning [Roich et al. 2022] and DreamBooth [Ruiz et al. 2023], we propose to finetune the parameters θ of VRMM with the prior preservation technique to better reproduce the individual characteristics:

$$\arg\min_{z_e,l,R,t,\phi,\theta} \mathcal{L}_{img} + \lambda_{LR}\mathcal{L}_{LR} + \lambda_{id}||z_{id}||_2^2 + \lambda_{exp}||z_e||_2^2, \tag{15}$$

where λ_{LR}, λ_{id}, and λ_{exp} are balancing weights. $\mathcal{L}_{LR}$ is the locality regularization term that restricts the modification to the local region around the inverted identity code z^*_{id}. Specifically, the interpolated identity code z^{inter}_{id} is obtained by linearly blending z^*_{id} with a randomly selected identity code z^i_{id} from the training set with a weight α:

$$z^{inter}_{id} = \alpha z^*_{id} + (1-\alpha) z^i_{id}. \tag{16}$$

Then we use the interpolated identity code z^{inter}_{id}, a randomly selected expression code z_e, and a lighting condition l to render an image with both the fine-tuned VRMM and the original VRMM. $\mathcal{L}_{LR}$ measures the difference of these two images similarly to $\mathcal{L}_{img}$.

We have found that fine-tuning for 500 iterations leads to good convergence, taking about five minutes. Our experiments show that the locality regularization significantly improves the prior preservation while having minor effect on the detail reconstruction.

4 EXPERIMENTS

4.1 Dataset

We capture dynamic facial performance of 254 subjects in a custom-built apparatus with synchronized multi-view cameras and controllable lighting condition. Each subject is asked to perform 21 predefined expressions, read out a paragraph, look at different directions and freely perform exaggerated and combined expressions.

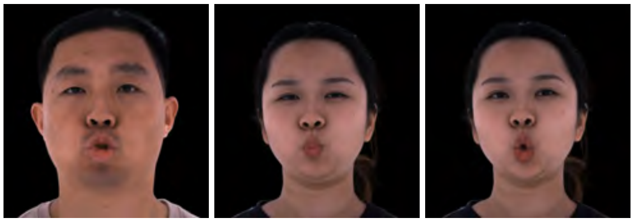

Figure 5: Ablation experiment about the expression consistency loss (ECL). By adding the expression consistency loss, our model learns a more unified expression space of different identities.

Figure 6: Ablation study on the locality regularization loss. The locality regularization loss helps our model to keep the knowledge from the prior during fine-tuning, resulting in accurate expression transfer after personalization.

We record 1800 frames for each subject at 16fps with 29 cameras surrounding the head with a resolution of 2448×2048. We utilize the group light pattern with basic background illumination [Bi et al. 2021; Yang et al. 2023] for relightable model capture. We also capture a full-on frame every four shots. The background images are recorded and alpha blended with the rendering during training similar to previous work [Lombardi et al. 2021; Yang et al. 2023]. The unprecedented dataset consists of more than 13M images in total with diverse dynamic expressions and known varying illuminations. Notice that the expressions do not need to be aligned or registered for different identities thanks to our self-supervised training framework, allowing for flexible dynamic performance capture. The network training on the dataset takes about two weeks on eight NVIDIA V100 graphics cards.

4.2 Qualitative Results

Latent space evaluation. Figure 8 shows interpolation between the identity codes of different individuals while keeping expression code and lighting condition fixed. The relighting results are shown in Figure 3. The visualization shows that the identity, expression and illumination spaces of our VRMM are fully disentangled and can be freely combined, which demonstrates the effectiveness of our self-supervised training framework.

Avatar personalization. Our model can be utilized to create personalized avatars from few-shot captures. Figure 9 shows that VRMM generates high-fidelity relightable and animatable avatars by fitting multi-view images from various sources. The first two rows are from the Multiface Dataset [Wuu et al. 2022], which are collected in a multi-view capture system with 40 and 146 cameras, respectively. We use the images of one frame as input. The middle two rows are from [Yang et al. 2023] with 23 input views. The last two rows are captured in our studio. The personalization is performed given only three images. Please refer to our accompanying video for the corresponding animations results.

Table 2: Quantitative evaluation results.

	Novel view synthesis			Single-view reconstruction		
Method	MAE ↓	SSIM ↑	LPIPS ↓	MAE ↓	SSIM ↑	LPIPS ↓
MoFaNeRF	28.3	0.829	0.247	35.00	0.878	0.199
HeadNeRF	17.19	0.892	0.211	14.07	0.905	0.163
Ours	**5.29**	**0.950**	**0.120**	**4.45**	**0.933**	**0.106**

4.3 Comparisons

We compare our method with existing publicly available parametric head models, *i.e.*, HeadNeRF [Hong et al. 2022] and MoFaNeRF [Zhuang et al. 2022].

Novel view synthesis. We conduct both qualitative and quantitative evaluations on the task of novel view synthesis using the Multiface Dataset [Wuu et al. 2022]. The parametric models are fitted to multi-view images from a single frame of each subject, and performance is measured against five held out views. The visual comparison is provided in Figure 4, with detailed quantitative results presented in Table 2.

Single-view reconstruction. Figure 7 shows qualitative comparison results for single-view head reconstruction. The average quantitative fitting error on 100 in-the-wild images from the FFHQ dataset [Karras et al. 2019] is reported in Table 2. These calculations are confined to the regions shared by different methods. Our method can achieve high-quality reconstruction results from a single in-the-wild image, while other baselines produce less accurate results.

4.4 Ablation Study

Expression consistency constraint. We conduct an ablation study to evaluate the effectiveness of the expression consistency loss (ECL). The experimental model is trained without expression consistency loss while keeping other parts identical. Then we set the expression code as the one corresponding to the target image to evaluate the expression consistency across identities. The results in Figure 5 show that our method learns more consistent expressions for different identities compared with the model w/o ECL.

Locality regularization loss. We attempt to remove our proposed locality regularization loss (LRL). The results are shown in Figure 6. We perform personalization using 39 images of a single frame from the Multiface Dataset. The personalized avatar is animated by the expression code corresponding to the target expression. The fine-tuning deteriorates the prior model without the locality regularization loss, which results in inaccurate animation results.

5 CONCLUSION

We present VRMM, the first volumetric morphable head model that enables continuous control over expression, identity, and global illumination. Through an elaborately designed training framework, VRMM is capable of disentangling complex attributes from multi-view sequences of casually varied expressions and lightings in a self-supervised manner. Once trained, VRMM can serve as a powerful prior for various reconstruction tasks. Combined with the novel fitting technique we propose, VRMM requires only minimal observations to accurately fit a specific portrait and generate an animatable and relightable avatar with real-time rendering. We believe this work will have a profound impact on the development of the field.

ACKNOWLEDGMENTS

We thank Xiaoqiang Liu for being our capture subject, the authors of [Zhuang et al. 2022] and [Hong et al. 2022] for releasing their source code and pretrained models, as well as anonymous reviewers for their insightful feedback.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. 2018. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3877–3886.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*. 1–9.

Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. 2011. High-quality passive facial performance capture using anchor frames. In *ACM SIGGRAPH 2011 papers*. 1–10.

Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn Mcphail, Ravi Ramamoorthi, Yaser Sheikh, and Jason Saragih. 2021. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–15.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. 2016. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5543–5552.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High resolution passive facial performance capture. In *ACM SIGGRAPH 2010 papers*. 1–10.

Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. 2023. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3402–3413.

Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.

Pol Caselles, Eduard Ramon, Jaime Garcia, Xavier Giro-i Nieto, Francesc Moreno-Noguer, and Gil Triginer. 2023. Sira: Relightable avatars from a single image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 775–784.

Zhaoxi Chen and Ziwei Liu. 2022. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*. Springer, 606–623.

Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. 2015. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–13.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.

Mingsong Dou, Philip Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. 2017a. Motion2fusion: Real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–16.

Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. 2017b. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5917.

Graham Fyffe and Paul Debevec. 2015. Single-shot reflectance measurement from polarized color gradient illumination. In *2015 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–10.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 1–10.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–19.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.

Haoda Huang, Jinxiang Chai, Xin Tong, and Hsiang-Tao Wu. 2011. Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition. In *ACM SIGGRAPH 2011 papers*. 1–10.

Zi-Hang Jiang, Qianyi Wu, Keyu Chen, and Juyong Zhang. 2019. Disentangled representation learning for 3D face shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11957–11966.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023).

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 760–769.

Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos Zafeiriou. 2021. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2021), 9269–9284.

Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. 2022. EyeNeRF: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–16.

Jiaman Li, Zhengfei Kuang, Yajie Zhao, Mingming He, Karl Bladin, and Hao Li. 2020b. Dynamic facial asset and rig generation from a single scan. *ACM Trans. Graph.* 39, 6 (2020), 215–1.

Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. 2020a. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3410–3419.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 194:1–194:17.

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–14.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 64–73.

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, Paul E Debevec, et al. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. *Rendering Techniques* 2007, 9 (2007), 10.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European Conference on Computer Vision*. 405–421.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. 2023. Relightify: Relightable 3D Faces from a Single Image via Diffusion

Models. *2023 IEEE/CVF International Conference on Computer Vision* (2023), 8772–8783. https://api.semanticscholar.org/CorpusID:258588229

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 165–174.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. Ieee, 296–301.

Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. 2020. Towards a complete 3D morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 11 (2020), 4142–4160.

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-shot high-quality facial geometry and skin appearance capture. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 1–12.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)* 42, 1 (2022), 1–13.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22500–22510.

Christophe Schlick. 1994. An inexpensive BRDF model for physically-based rendering. In *Computer graphics forum*. 233–246.

Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. 2020. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 91:1–91:15.

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1274–1283.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2387–2395.

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–14.

Luan Tran and Xiaoming Liu. 2018. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7346–7355.

Luan Tran and Xiaoming Liu. 2019. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1 (2019), 157–171.

Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face Transfer with Multilinear Models. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 426–433.

Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. 2022. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. 2022. Multiface: A Dataset for Neural Face Rendering. *arXiv preprint arXiv:2207.11243* (2022).

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. 2023. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–10.

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Haotian Yang, Mingwu Zheng, Wanquan Feng, Haibin Huang, Yu-Kun Lai, Pengfei Wan, Zhongyuan Wang, and Chongyang Ma. 2023. Towards Practical Capture of High-Fidelity Relightable Avatars. In *SIGGRAPH Asia 2023 Conference Proceedings*.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. Facescape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 601–610.

Longwen Zhang, Chuxiao Zeng, Qixuan Zhang, Hongyang Lin, Ruixiang Cao, Wei Yang, Lan Xu, and Jingyi Yu. 2022. Video-driven neural physically-based facial asset for production. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

Mingwu Zheng, Zhang Haiyu, Hongyu Yang, and Di Huang. 2023a. NeuFace: Realistic 3D Neural Face Rendering from Multi-view Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 16868–16877.

Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. 2022. ImFace: A Nonlinear 3D Morphable Face Model with Implicit Neural Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 20343–20352.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023b. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21057–21067.

Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. 2020. Improved stylegan embedding: Where are the good latents? *arXiv preprint arXiv:2012.09036* (2020).

Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. 2017. Face alignment in full pose range: A 3D total solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 1 (2017), 78–92.

Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. 2022. Mofanerf: Morphable facial neural radiance field. In *European Conference on Computer Vision*. Springer, 268–285.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4574–4584.

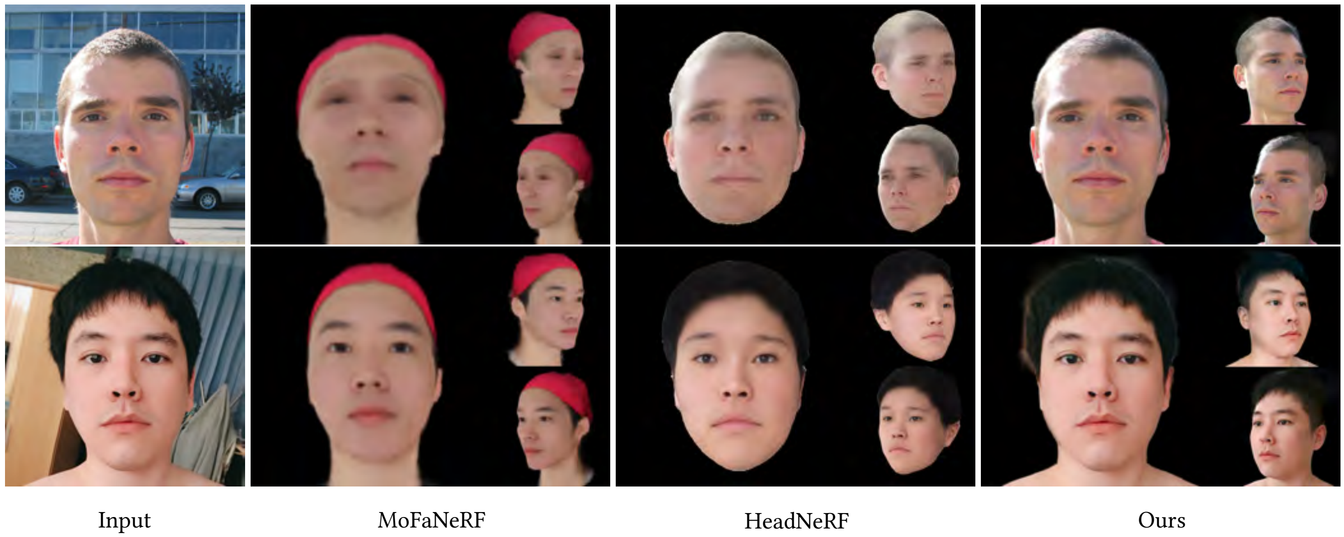

Figure 7: Comparison with existing volumetric head priors, MoFaNeRF [Zhuang et al. 2022] and HeadNeRF [Hong et al. 2022], on single-view head reconstruction. We show the fitting results in the original view as well as two different views. The reconstruction results of our method achieve significantly better visual quality.

Figure 8: Interpolation results between three identities (left, center, right). Our model learns a smooth identity latent space that allows linear interpolation. Besides, the expression keeps unchanged during the interpolation, confirming that the expression and identity spaces have been effectively disentangled.

Figure 9: Few-shot personalization results. Given multi-view images of a subject in a static expression under unknown fixed illumination, our model enables the creation of a high-fidelity avatar that can be animated and relighted. Our model learns the relightable appearance even when the input images suffer from strong specular reflection.