

Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures

<https://syntec-research.github.io/Cafca>

MARCEL C. BUEHLER, ETH Zurich, Switzerland
 GENGYAN LI, ETH Zurich and Google, Switzerland
 ERROLL WOOD, Google, UK
 LEONHARD HELMINGER, Google, Switzerland
 XU CHEN, Google, Switzerland
 TANMAY SHAH, Google, USA
 DAOYE WANG, Google, Switzerland
 STEPHAN GARBIN, Google, UK
 SERGIO ORTS-ESCOLANO, Google, Switzerland
 OTMAR HILLIGES, ETH Zurich, Switzerland
 DMITRY LAGUN, Google, USA
 JÉRÉMY RIVIERE, Google, Switzerland
 PAULO GOTARDO, Google, Switzerland
 THABO BEELER, Google, Switzerland
 ABHIMITRA MEKA, Google, USA
 KRIPASINDHU SARKAR, Google, Switzerland

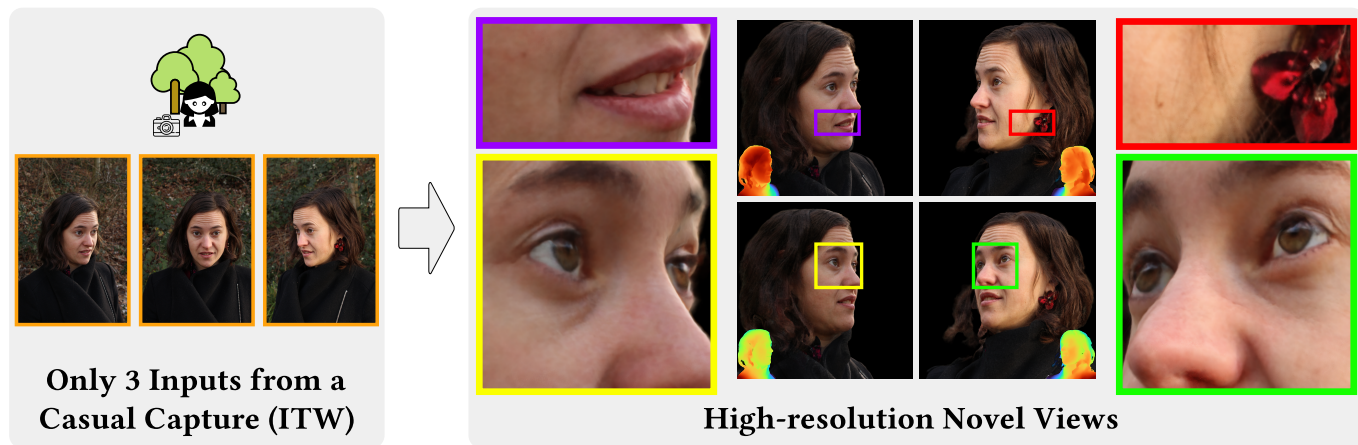


Fig. 1. We propose a method for high-quality novel view synthesis of expressive faces captured in-the-wild. From a casual capture with only three images, our method synthesizes novel views at an unprecedented level of detail showing wrinkles, hair strands, skin pores, and eyelashes.

Volumetric modeling and neural radiance field representations have revolutionized 3D face capture and photorealistic novel view synthesis. However, these methods often require hundreds of multi-view input images and are

thus inapplicable to cases with less than a handful of inputs. We present a novel volumetric prior on human faces that allows for high-fidelity expressive face modeling from as few as three input views captured in the wild. Our key insight is that an implicit prior trained on synthetic data alone can generalize to extremely challenging real-world identities and expressions and render novel views with fine idiosyncratic details like wrinkles and eyelashes. We leverage a 3D Morphable Face Model to synthesize a large training set, rendering each identity with different expressions, hair, clothing, and other assets. We then train a conditional Neural Radiance Field prior on this synthetic dataset and, at inference time, fine-tune the model on a very sparse set of real images of a single subject. On average, the fine-tuning

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1131-2/24/12.

<https://doi.org/10.1145/3680528.3687580>

requires only three inputs to cross the synthetic-to-real domain gap. The resulting personalized 3D model reconstructs strong idiosyncratic facial expressions and outperforms the state-of-the-art in high-quality novel view synthesis of faces from sparse inputs in terms of perceptual and photo-metric quality.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; *Shape representations*; *Appearance and texture representations*; Neural networks; **Volumetric models**.

Additional Key Words and Phrases: Neural rendering, face reconstruction, novel view synthesis, sparse reconstruction, synthetic data

ACM Reference Format:

Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, J  r  my Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarker. 2024. Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures: <https://syntec-research.github.io/Cafca>. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24)*, December 3–6, 2024, Tokyo, Japan. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3680528.3687580>

1 INTRODUCTION

“Who sees the human face correctly: the photographer, the mirror, or the painter?” - Pablo Picasso. Our visual acuity is remarkably hyper-tuned to perceive the details of human faces due to evolutionary design [Pascalis and Kelly 2009]. Individual-specific expressions play a particularly significant role in perception tasks such as identification or estimation of emotion and intent [Kret 2015; Lee and Anderson 2017; Sinha et al. 2006]. Highly personalizable 3D representations that model the idiosyncracies of face shape and deformation at high quality are crucial to truly immersive and photorealistic 3D portrait photography. Perhaps, if trained well, an AI could join Picasso’s list.

Wide-scale adoption and democratization of 3D portrait photography demands casual captures in uncontrolled environments – *i.e.*, only a few shots taken with a handheld camera. Volumetric representations [Barron et al. 2022; Mildenhall et al. 2020; Park et al. 2021a,b], have demonstrated impressive quality and photorealism in synthesizing novel views of a large variety of 3D scenes using densely captured 2D images. Extended works relax the requirement for dense capture through various forms of regularization – for instance, entropy minimization [Rebain et al. 2022], spatial smoothness [Niemeyer et al. 2022], depth regularization [Guangcong et al. 2023; Prinzler et al. 2023] and volume bounds [Sarker et al. 2023]. These methods target general scenes and still require dozens of views captured simultaneously. As such, they cannot be applied to expressive 3D portrait photography in the wild, due to ambiguities that arise from the limited input and uncontrolled conditions.

Lifting very sparse 2D views onto a 3D reconstruction requires a strong face prior, crafted using diverse data captured across the human population. However, large real datasets are expensive and challenging to collect and typically suffer from low resolution, sampling limitations, and biases in their coverage of diversity and detail of facial geometry and appearance, under varied viewpoints and lighting. Here, our key insight is that such prior can be built from synthetic data alone and fine-tuned on a few real-world images to bridge the synthetic-real domain gap, generalizing robustly to challenging portraits captured in the wild, as shown in Figs. 1 and 9.

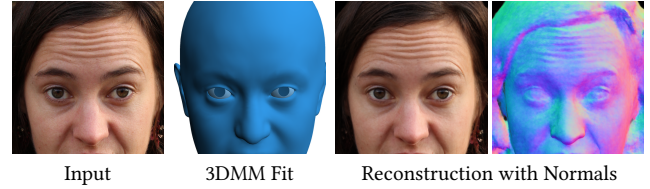


Fig. 2. Our method reconstructs details that go far beyond the capabilities of our synthetic dataset. We show a crop of one of the input views from the teaser and its corresponding 3DMM fit. While the 3DMM lacks details, our reconstruction models the wrinkles on the forehead.

Recent works have shown that well-curated, calibrated, and diverse datasets can be built from synthetic graphical renderings. Such data has been successfully applied to various face perception tasks, including sparse problems like landmark localization and segmentation [Wood et al. 2021, 2022] and dense problems like view synthesis [Sun et al. 2021] and relighting [Yeh et al. 2022]. However, such synthetic data is unable to model the full light transport of complex interactions like sub-surface scattering, specular polarization, and global illumination, thus presenting a significant domain gap. Interestingly, sparse regression problems can robustly overcome such gaps at inference time. In contrast, denser synthesis problems cannot, requiring domain adaptation solutions that suffer from quality losses that prevent their use in building a 3D face prior.

We propose a novel, volumetric facial prior that is learned from a large dataset of synthetic images of different identities and expressions, rendered with accessories like hair and beard styles, glasses, clothing, and skin textures. At inference time, given a few input images at arbitrary resolutions, we fine-tune this prior to reconstruct a high-quality, personalized volumetric model of the captured subject. This new model then allows for 3D-consistent novel view synthesis at high resolution, even when the original input shows exaggerated facial expressions and challenging lighting conditions. We overcome the domain gap between synthetic and real by using an MLP-based generator for volumetric rendering. This MLP learns a strong prior over geometry while generalizing to unseen appearance domains, including colored lighting and shadows. We demonstrate the efficacy of our method through extensive evaluation, ablations, and comparison with the state of the art.

In summary, our key contributions are:

- We show that synthetic data alone can be successfully leveraged to learn a strong face prior for few-shot, personalized 3D face modeling in the wild.
- Our new model and few-shot fine-tuning method can robustly reconstruct expressive faces under challenging in-the-wild lighting and synthesize photorealistic novel views with unprecedented quality and fine-scale idiosyncratic details.

Our synthetic dataset is available for research purposes at <https://syntec-research.github.io/Cafca>.

2 RELATED WORK

Multiple works have explored ways to mitigate the data-intensive nature of volumetric reconstruction, where some main themes have

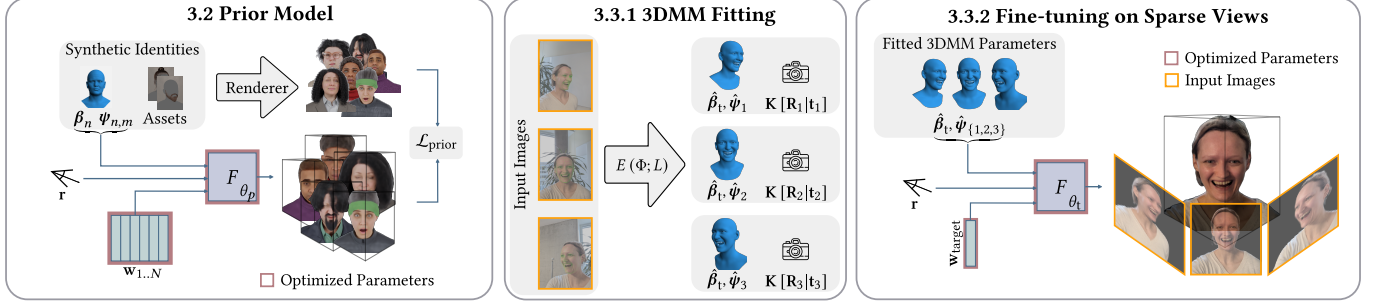


Fig. 3. Method Overview. We train an implicit prior model on renderings of a 3DMM combined with assets like hair, beard, clothing, and more (Sec. 3.2). For inference, we first estimate the camera, identity, and expression parameters (Sec. 3.3.1) and then fine-tune the implicit model (Sec. 3.3.2). The implicit prior trained on synthetic data alone can generalize to extremely difficult real-world expressions and render fine details like wrinkles and eyelashes (Figs. 1, 9). While we show three input views for the average case, our method can also work with one (Fig. 10) or more input views (supplementary material).

emerged: regularization schemes [Jain et al. 2021; Niemeyer et al. 2022; Rebain et al. 2022; Truong et al. 2023; Yang et al. 2023], carefully crafted initialization of model parameters [Kundu et al. 2022; Tancik et al. 2021; Vora et al. 2021], depth signals or feature embeddings [Guangcong et al. 2023; Jain et al. 2021; Truong et al. 2023], and data-driven, pre-trained priors [Buehler et al. 2023; Chan et al. 2022; Chen et al. 2021; Gu et al. 2021; Jain et al. 2021; Mihajlovic et al. 2022; Or-El et al. 2022; Prinzler et al. 2023; Ramon et al. 2021; Rebain et al. 2022; Tan et al. 2022; Truong et al. 2023; Wang et al. 2021; Yu et al. 2021; Zhang et al. 2022].

Regularization. RegNeRF [Niemeyer et al. 2022] employs a smoothness regularization term on the expected depth derived from the predicted density, in conjunction with a patch-based regularizer on appearance which together bias the model towards a 3D consistent solution. FreeNeRF [Yang et al. 2023] proposes a gradual, coarse-to-fine training scheme to prevent overfitting caused by high-frequency, position encoding components early in the fit. Despite their promising results on in-the-wild images, these methods cannot yet provide high-quality reconstructions from few-shot inputs, as we show in Fig. 7.

Initialization. A common strategy is to learn initial model parameters [Finn et al. 2017; Nichol et al. 2018; Sitzmann et al. 2020; Tancik et al. 2021; Zakharov et al. 2019] from a large collection of images. While this strategy has been shown to offer faster convergence, its applicability to high-resolution settings remains challenging due to computational requirements of large neural networks.

Data-driven Priors. With the advent of large datasets of particular domains [Choi et al. 2020; Karras et al. 2019; Liu et al. 2018], recent works have explored data-driven priors trained on a corpus of data specific to the reconstruction task at hand. Generative neural field models in particular [Chan et al. 2022, 2021; Deng et al. 2022a,b; Gu et al. 2021; Niemeyer and Geiger 2021; Rao et al. 2022; Rebain et al. 2022; Schwarz et al. 2020; Tan et al. 2022; Wang et al. 2022; Zhou et al. 2021] have shown promising results. To tackle the heavy computational and memory requirement of volumetric rendering, EG3D [Chan et al. 2022] proposes to use a lightweight tri-plane feature representation. Most of these models further rely on an extra 2D super-resolution step [Chan et al. 2022, 2021; Gu et al.

2021; Hong et al. 2022; Tan et al. 2022; Wang et al. 2023] as they can only be trained at low resolutions due to high memory footprint requirements.

Recent works [Buehler et al. 2023; Cao et al. 2022; Wang et al. 2022] have overcome the need for 2D super-resolution. MoRF [Wang et al. 2022] learns a conditional neural reflectance field of faces from a small training dataset comprising of 12 views of 15 real subjects making neutral expressions, further augmented with synthetic renderings. [Cao et al. 2022] trains an avatar prior with a Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021] from data captured in a controlled environment. They leverage their model to fine-tune to a specific target subject’s identity based on a short sequence of a casually captured RGB-D video. Another line of work leverages 2D generative priors [Liu et al. 2023; Massague et al. 2024; Melas-Kyriazi et al. 2023; Tang et al. 2023; Wu et al. 2024; Zeng et al. 2023]. These methods typically employ score distillation sampling [Poole et al. 2022; Wang et al. 2024] for injecting signals from unseen views given text prompts or sparse input images.

Multiple recent works have leveraged 3D Morphable Face Models (3DMM) [Blaiz and Vetter 1999] as strong priors. This has led to the development of avatars that can be driven by 3DMM expression coefficients [Buehler et al. 2021; Duan et al. 2023; Li et al. 2024; Niemeyer et al. 2022; Xu et al. 2023; Zhao et al. 2023; Zheng et al. 2022, 2023]. Such avatars are typically trained on a sequence of monocular video frames showing various head poses and facial expressions. Other works have explored the reconstruction of faces and textures from a single input image [Gao et al. 2020; Papantoniou et al. 2023; Vinod et al. 2024]. Another body of work has explored novel view synthesis from sparse inputs by training their model as an auto-encoder, coupled with image-based rendering. These methods [Chen et al. 2021; Mihajlovic et al. 2022; Wang et al. 2021; Yu et al. 2021] typically train a convolutional encoder that maps input images to 2D feature maps in images-space to condition the scene’s volume. This approach can typically be extended with additional priors such as keypoints [Mihajlovic et al. 2022], depth [Guangcong et al. 2023; Prinzler et al. 2023; Xu et al. 2022] or pixel-matches across input views [Truong et al. 2023].

Preface [Buehler et al. 2023] is a method for high-quality face avatar creation in which they train a low-resolution, generative

prior on a dataset of facial identities captured in-studio. They could not reconstruct strong expressions due to the domain limitation of their prior model trained only on neutral faces. Our method tackles this challenging problem and we demonstrate compelling examples of novel view synthesis of expressive faces from sparse inputs.

3 METHOD

We present a method that takes as input as few as three images of a person’s face – of arbitrary identity, expression, and lighting condition – and reconstructs a personalized 3D face model that can render high-quality, photorealistic novel views of that person, including fine details like freckles, wrinkles, eyelashes, and teeth.

To overcome reconstruction ambiguities, our method uses a pre-trained volumetric face model as prior, trained on a large dataset of synthetic faces with a variety of identities, expressions, and viewpoints rendered in a single environment. At inference time, our method first fits the coefficients of our prior model to a small set of real input images. It then further fine-tunes the model weights and effectively performs domain adaptation during the few-shot reconstruction process, see Fig. 3. While the prior model is trained only once on a large collection of synthetic face images, the inference-time optimization is performed on a per-subject basis from as few as three (*e.g.*, smartphone) images captured in-the-wild (see Fig. 9).

This section starts with background information (Sec. 3.1), details the training of synthetic prior (Sec. 3.2) and then finetuning from three input views (Sec. 3.3).

3.1 Background: NeRF and Preface

NeRF. Our prior face model is built upon Neural Radiance Fields (NeRF) [Mildenhall et al. 2020]. NeRFs represent 3D objects as density and (emissive) radiance fields parameterized by neural networks. Given a camera ray \mathbf{r} , a NeRF samples 3D points \mathbf{x} along the ray that are fed together with the view direction \mathbf{d} into an MLP. The output is the corresponding density σ and color value \mathbf{c} at \mathbf{x} . A NeRF is rendered into any view via volumetric rendering. The color $\mathbf{c}(\mathbf{p})$ of a pixel \mathbf{p} is determined by compositing the density and color along the camera ray \mathbf{r} within an interval between a near and a far camera plane $[t_n, t_f]$:

$$\mathbf{c}(\mathbf{p}) = \mathbf{F}_\theta(\mathbf{r}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t)) \mathbf{c}(\mathbf{r}(t), \mathbf{d}) dt, \quad (1)$$

$$\text{where } T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s)) ds\right). \quad (2)$$

The variable θ denotes the model parameters that are fit to the input data.

Preface. Our method also extends Preface [Buehler et al. 2023], a method for novel view synthesis of neutral faces given sparse inputs. Besides the position \mathbf{x} and view direction \mathbf{d} , Preface also takes a learned latent code \mathbf{w} as input. The latent code represents the identity and is optimized while training the model as an auto-decoder [Bojanowski et al. 2018]. During inference, Preface first projects the sparse input images into its latent space, by optimizing one identity code \mathbf{w} . Then, it fine-tunes all model parameters under regularization constraints on the predicted normals and the

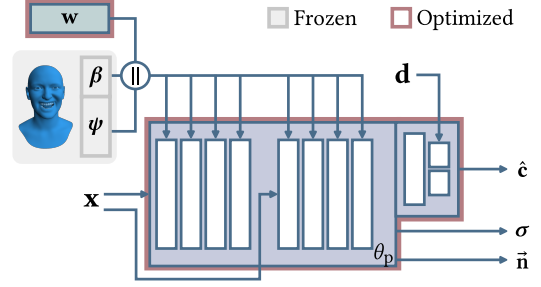


Fig. 4. We implement our prior model as a conditional NeRF [Barron et al. 2022; Mildenhall et al. 2020]. We condition by concatenating (\parallel) three codes: a 3DMM identity code β , a 3DMM expression code ψ , and a learned latent code \mathbf{w} representing out-of-model characteristics like hair, clothing, etc. The outputs include the color $\hat{\mathbf{c}}$, the density σ , and a normal vector $\hat{\mathbf{n}}$. Please see Sec. 2 in the supp. PDF for details.

view weights. While Preface excels at high-resolution novel view synthesis of neutral faces, it struggles in the presence of strong, idiosyncratic expressions (Fig. 7). In the following, we address this limitation while building an improved prior from synthetic images alone.

3.2 Pretraining an Expressive Prior Model

We train a prior model to capture the distribution of human heads with arbitrary facial expressions. As Preface [Buehler et al. 2023], our prior model is implemented as a conditional Neural Radiance Field (NeRF) [Mildenhall et al. 2020; Rebain et al. 2022] with a Mip-NeRF 360 backbone [Barron et al. 2022], Fig. 4. Yet, we observe that the simple Preface auto-decoder [Bojanowski et al. 2018; Rebain et al. 2022] cannot achieve high-quality fitting results on expressive faces (see the ablation in Sec. 4.3). We hypothesize that the distribution of expressive faces is much more difficult to model and disentangle than the distribution of neutral faces. To address this limitation, we decompose the latent space of our prior model into three latent spaces: a predefined identity space \mathcal{B} , a predefined expression space Ψ , and a learned latent space \mathcal{W} . The identity and expression spaces come from a linear 3D Morphable Model (3DMM) in the style of [Wood et al. 2021]. The latent codes in these two spaces are known *a priori* and represent the face shape for the arbitrary identity and expression in each synthetic training image. These codes are also frozen and do not change during training. The latent space \mathcal{W} represents characteristics that are not modeled by the 3DMM like hair, beard, clothing, glasses, appearance, lighting, etc., and is learned while training the auto-decoder as in Preface [Buehler et al. 2023]. Considering this model, we adapt Eq. 1 to obtain:

$$\mathbf{F}_{\theta_p}(\mathbf{r}, \beta, \psi, \mathbf{w}) = \int_{t_n}^{t_f} T(t) \sigma(\mathbf{r}(t), \beta, \psi, \mathbf{w}) \mathbf{c}(\mathbf{r}(t), \mathbf{d}, \beta, \psi, \mathbf{w}) dt, \quad (3)$$

$$\text{where } T(t) = \exp\left(-\int_{t_n}^t \sigma(\mathbf{r}(s), \beta, \psi, \mathbf{w}) ds\right), \quad (4)$$



Fig. 5. Our synthetic faces exhibit a wide range of geometric and expression diversity and can be rendered from any viewpoint.

where $\beta \in \mathcal{B} \subset \mathbb{R}^{48}$ and $\psi \in \Psi \subset \mathbb{R}^{157}$ are the 3DMM identity and expression parameters, and $\mathbf{w} \in \mathcal{W} \subset \mathbb{R}^{64}$ is a learned parameter encoding additional characteristics.

We train this prior on synthetic data alone – it never sees a real face. While it would be feasible to train a prior model on real data (see our ablation in Tbl. 2), we chose synthetic over real for multiple reasons. Real datasets exhibit limited diversity. Most face datasets feature monocular frontal views only, with few expressions other than smiles. Some multi-view, multi-expression datasets exist [Kirschstein et al. 2023; Wu et al. 2022; Zhu et al. 2023], but consist of relatively few individuals due to the complexity and expense of running a capture studio. Further, subjects must adhere to wardrobe restrictions: glasses are forbidden and hair must be tucked away. A prior trained on such data will not generalize well to expressive faces captured in the wild. Besides, the logistics of capturing large-scale real data is extremely expensive, time and energy-consuming, and cumbersome. Instead, synthetics guarantee us a wide range of identity, expression, and appearance diversity, at orders of magnitude lower cost and effort. In addition, synthetics provide perfect ground truth annotations: each render is accompanied by 3DMM latent codes β, ψ .

We synthesize facial training data as in Wood et al. [2021]. We first generate the 3D face geometry by sampling the identity and expression spaces of the 3DMM. We then make these faces look realistic by applying physically based skin materials, attaching strand-based hairstyles, and dressing them up with clothes and glasses from our digital wardrobe. The scene is rendered with environment lighting using Cycles, a physically-based ray tracer (www.cycles-renderer.org). Examples are shown in Fig. 5 and on the supplementary HTML page. To help disentangle identity from expression, we sample 13 different random expressions for each random identity. Each expression is then rendered from 30 random viewpoints around the head. All faces are rendered under the same lighting condition, which was chosen to minimize shadows on the face.

Each training iteration randomly samples rays \mathbf{r} from a subset of all identities and expressions. A ray is rendered into a pixel color as given by Eq. 3. We optimize the network parameters θ_p and \mathbf{N}

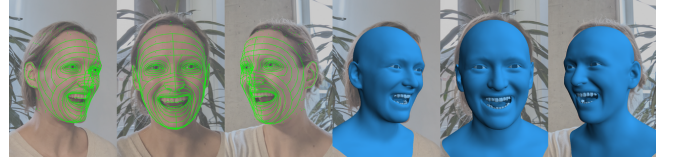


Fig. 6. For inference, we first recover the 3DMM and camera parameters through model-fitting to probabilistic 2D landmarks (Sec. 3.3.1).

per-identity latent codes $\mathbf{w}_{1..N}$ while keeping the 3DMM expression and identity codes β and ψ frozen:

$$\theta_p, \mathbf{w}_{1..N} = \arg \min_{\theta, \mathbf{w}_{1..N}} \mathcal{L}_{\text{prior}}, \quad \mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{prop}} \mathcal{L}_{\text{prop}}. \quad (5)$$

Here $\mathcal{L}_{\text{recon}}$ is the mean-absolute error between the predicted and ground-truth colors, and $\mathcal{L}_{\text{prop}}$ is the weight distribution matching loss from Mip-NeRF [Barron et al. 2021]. Please see Sec. 2 in the supp. PDF for the spelled-out loss terms.

We find that, when training from scratch, the model quickly collapses and outputs zero densities everywhere. We solve this by first training on images with background for 50,000 steps and continuing without background.

3.3 Inference from Sparse Views

We use our low-resolution synthetic prior model to enable high-resolution novel view synthesis of real expressive faces from few input images. We first describe how we obtain the conditional inputs and camera parameters in Sec. 3.3.1 and the subsequent fine-tuning of our model in Sec. 3.3.2.

3.3.1 3DMM Fitting and Camera Estimation. Figure 3 gives an overview of the 3DMM fitting. During inference, the first step is to recover camera and 3DMM parameters from un-calibrated input images. We follow the approach of previous work [Wood et al. 2022] and fit to dense 2D landmarks (see Fig. 6). We first predict 599 2D probabilistic landmarks. Each landmark corresponds to a vertex in our 3DMM and is predicted as a 2D isotropic Gaussian with expected 2D location μ and scalar uncertainty σ . Next, we minimize an energy $E(\Phi; L)$, where L denotes the landmarks and Φ all the optimized 3DMM parameters including identity, expressions, joint rotations, and global translation, and intrinsic and extrinsic camera parameters, if unknown.

Minimizing E encourages the 3DMM to explain the observed landmarks with a probabilistic 2D landmark energy, and discourages unlikely faces using regularizers on 3DMM parameters and mesh self-intersection (additional detail is given in [Wood et al. 2022] and in Sec. 2 of the supp. mat.).

The benefit of the 3DMM fitting is two-fold. First, we get a good estimate of the world position of the camera and the head, so that later during inversion and finetuning of the model, camera parameters can be frozen. Second, thanks to the alignment of the 3DMM latent space and our prior model’s latent space, we directly feed the 3DMM parameters into the model, which serves as a good initialization during the subsequent inversion stage.

The outputs of this step are the camera parameters (shared intrinsics \mathbf{K} and per-camera extrinsics $[\mathbf{R}_i | \mathbf{t}_i]$), a shared identity code β ,

and per-image expression codes ψ_i . For casual in-the-wild captures, it can be challenging to hold the same expression while the data is being captured. Therefore, we allow expression code to vary slightly between images to make the inversion robust to small, involuntary micro-changes in expression. In the studio setting, however, the cameras are synchronized and hence it is sufficient to optimize for a single, shared expression code.

3.3.2 Fine-tuning on Sparse Views. This section describes how to fine-tune the low-resolution synthetic prior model to sparse, high-resolution real input images. Fine-tuning requires a short warm-up phase where only the latent code for the target $\mathbf{w}_{\text{target}}$ is optimized. After that, fine-tuning optimizes all model parameters under additional constraints on the geometry and the appearance weights. We randomly sample rays from all available inputs, typically three images, and mask them to the foreground by multiplying them by an estimated foreground mask [Pandey et al. 2021].

Warm-up by Latent Code Inversion. While the 3DMM fitting provides the identity and expression codes β, ψ_i , our model also requires the conditional input \mathbf{w} , which models out-of-model characteristics like hair, clothing, and appearance. We follow [Buehler et al. 2023] and search the learned latent space \mathcal{W} of the prior model for a latent code that roughly matches the geometry and appearance of the input images. We downscale the three input images to the resolution of the prior model, sample random patches, and optimize

$$\mathbf{w}_{\text{target}} = \arg \min_{\mathbf{w}} \mathcal{L}_{\text{recon}} + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}. \quad (6)$$

The photo-metric reconstruction term $\mathcal{L}_{\text{recon}}$ is the mean absolute error between the rendered and the ground-truth patch and LPIPS is a perceptual loss in the feature space of a pre-trained image classifier [Simonyan and Zisserman 2015; Zhang et al. 2018]. Note that the LPIPS loss is only employed during inversion, not during model fitting. The camera, 3DMM identity, and expression parameters are frozen during inversion.

Model Fitting. The output of the warm-up is a rough approximation of the input images in a *low-resolution, synthetic* space. In model fitting, we cross the domain gap to enable detailed novel view synthesis at *high resolution* for *realistic* faces. The model fitting needs to cross a substantial domain gap so that the output can contain details that have never been seen during prior model training.

Model fitting optimizes all model parameters on sparse, usually two or three, input images. In a randomly initialized NeRF [Mildenhall et al. 2020], this optimization would overfit and would fail to produce correct novel views [Buehler et al. 2023; Truong et al. 2023; Yang et al. 2023]. Thanks to our pretrained prior model, we can employ both *implicit* and *explicit* regularization, which yields high-quality results even in such sparse settings.

Implicit regularization comes from the fact that our prior model is trained on an aligned dataset of human faces. Initializing the weights of a NeRF with the correct latent code and weights of a prior model avoids total collapse. However, the optimization can still produce strong artifacts like duplicate ears and view-dependent color distortions. We follow [Buehler et al. 2023] and add explicit regularization on the consistency of predicted vs. analytical normals

Table 1. Quantitative evaluation on the Multiface dataset [Wuu et al. 2022].

Method	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Sparse NeRF [Guangcong et al. 2023]	16.29	0.6470	0.4024
Diner [Prinzler et al. 2023]	17.08	0.6608	0.3123
SPARF [Truong et al. 2023]	18.38	0.6401	0.4032
FreeNeRF [Yang et al. 2023]	21.42	0.7093	0.3612
Preface [Buehler et al. 2023]	25.02	0.7539	0.3129
Ours	26.49	0.7721	0.2970

and an L2 regularization on the weight of the view branch to avoid view-dependent flickering. In addition, we add a distortion loss term [Barron et al. 2022] $\mathcal{L}_{\text{dist}}$ for a more compact geometry:

$$\theta_t, \mathbf{w}_{\text{target}} = \arg \min_{\theta, \mathbf{w}} \mathcal{L}_{\text{fit}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{prop}} \mathcal{L}_{\text{prop}} + \lambda_{\text{normal}} \mathcal{L}_{\text{normal}} + \lambda_d \mathcal{L}_d + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}} \quad (7)$$

where $\mathcal{L}_{\text{normal}}$ and \mathcal{L}_d are the regularizers on predicted normals and view weights from Preface. We list and explain all loss terms in more detail in Sec. 2 of the supp. PDF.

Inference In-the-wild. For in-the-wild (ITW) images, we capture three images sequentially with a hand-held camera. The captured face might inhibit small movements during the capture, called micromotions. To mitigate these micromotions, we fine-tune with individual expression code for every input image. The 3DMM fitting yields a per-image expression code $\hat{\psi}_i$. During inference, we interpolate these expression codes based on their distance to the target camera. The weight is computed as the inverse squared distance between the target and all training cameras. The interpolated expression code $\tilde{\psi}_t$ for a target camera is computed as

$$\tilde{\psi}_t = \sum_i \underbrace{\frac{Z}{\epsilon + \|\tilde{\mathbf{l}}_t - \hat{\mathbf{l}}_i\|_2^2}}_{\text{weight}} \underbrace{\hat{\psi}_i}_{\text{training expression}}, \quad (8)$$

where $\hat{\psi}_i$ are the expression codes of the training frames, $\tilde{\mathbf{l}}_t$ is the position of the target camera, $\hat{\mathbf{l}}_i$ are the positions of the training cameras, ϵ is a small constant, and Z is a normalization factor to ensure that the weights sum up to 1.

4 EXPERIMENTS

This section presents an extensive evaluation of our proposed method with both quantitative and qualitative comparisons to related work and additional ablations. For comparisons, we run publicly available code for SparseNeRF [Guangcong et al. 2023], Sparf [Truong et al. 2023], and Diner [Prinzler et al. 2023]. For FreeNeRF [Yang et al. 2023] and Preface [Buehler et al. 2023], we use our own implementation. We compare renders at the resolution 1334×2048 pixels, except for Diner, where we render at a lower resolution (160×256 pixels) due to its architecture and memory constraints. Please refer to the supplementary video and HTML page for more results.

Table 2. We ablate the performance for different variants of the prior model. We pre-train a) on a smaller training set with fewer subjects, b) for a shorter number of steps, c) with different configurations of our synthetic rendering pipeline, and d) include real data. The *full* prior model is trained on 1,500 subjects with 13 expressions each for 1 Mio. steps on synthetic renderings with all available accessories in a single environment map. Please see the supp. mat. for details of the rendering pipeline.

	Pre-training Variant	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
a.i)	No Pre-training	10.21	0.3448	0.4256
a.ii)	On 1 subject	24.00	0.7512	0.3358
a.iii)	On 15 subjects	25.47	0.7668	0.3248
a.iv)	On 1,500 subjects	26.54	0.7750	0.3144
b.i)	For 105K steps	26.64	0.7752	0.3208
b.ii)	For 500K steps	26.54	0.7750	0.3144
b.iii)	For 1 Mio. steps	26.49	0.7721	0.2970
c.i)	On gray-scale renderings	26.44	0.7740	0.3242
c.ii)	On low-quality renderings	26.53	0.7727	0.3401
c.iii)	In diverse environments	26.58	0.7727	0.3415
c.iv)	Without makeup	26.72	0.7755	0.3210
c.v)	Without accessories	26.00	0.7731	0.3484
c.vi)	Without hair	26.14	0.7727	0.3303
c.vii)	With all accessories	26.54	0.7750	0.3144
d.i)	On real images	26.14	0.7708	0.3237
d.ii)	On real <i>and</i> synthetic images	26.41	0.7726	0.3227
d.iii)	On synthetic images	26.54	0.7750	0.3144
e)	Full	26.49	0.7721	0.2970

4.1 Quantitative Evaluation

Performance on casual in-the-wild captures is inherently difficult to evaluate quantitatively due to the lack of proper validation views. Often, the captured subject can hardly remain completely still. Therefore, we quantitatively evaluate on the Multiface [Wuu et al. 2022] studio dataset on nine scenes by randomly selecting three subjects with three expressions per subject. For each test subject, we use one frontal and two side views as input for training, as shown in Fig. 7. We remove the background by multiplying a foreground alpha matte estimated by [Pandey et al. 2021]. We hold out from 26 to 29 validation images per subject, where the camera viewing direction is located in the frontal hemisphere (see details in the supplementary material). As evaluation metrics, we measure photo-metric distance and image similarity via PSNR, SSIM, and LPIPS [Zhang et al. 2018], as summarized in Tbl. 1. Note that photometric reconstruction metrics like PSNR and SSIM and perceptual metrics like LPIPS are at odds with each other [Blau and Michaeli 2018]. We handle this tradeoff by optimizing perceptual quality in the warm-up (Eq. 6) and reconstruction quality in the fine-tuning (Eq. 7).

As shown in Tbl. 1, our new method provides better modeling fidelity across all three metrics over the set of validation views. In particular, compared to the state-of-the-art Preface [Buehler et al. 2023], we achieve the following improvement on the computed metrics: 6% PSNR, 2.4% SSIM, and 5% LPIPS. Visually, Fig. 7 shows that the novel views generated by our fine-tuned model more closely resemble the facial shape and appearance, including eye and mouth details, of the example (ground-truth) validation view.

4.2 Qualitative Results in the Wild

To demonstrate the robustness of our method in the wild, we captured subjects with a handheld DSLR camera (Canon EOS 4000D) and mobile phones (Pixel 7 and Pixel 8 Pro). We capture between one and three images per subject in various outdoor and indoor environments, including very challenging lighting conditions. These diverse in-the-wild results are shown in Fig. 1, Fig. 8, Fig. 9, and in the supplementary video and HTML page. The inlays in Fig. 1 show the high level of modeled detail on the lips and on the individual hair strands and eyelashes. Fig. 8 compares with the best performing related work [Buehler et al. 2023], which struggles with the mouth region. Of particular note is also the “tongue-out” expression in the third row of Fig. 9. Note that our synthetic training data contains tongues but the tongues never stick out of the mouth.

We also consider the more challenging single input image scenario, see Fig. 10. Our method achieves high-fidelity synthesis of the frontal face even for stylized and painted faces (right) but quickly degrades for side views. This behavior is expected, as our model is only trained on low-resolution synthetic images and has never seen high-resolution views from the side of a face. Please see the supp. HTML page for more examples.

Additional qualitative comparisons to previous work are shown in Fig. 7. Our method outperforms the other baseline methods in two main ways. First, our approach captures fine identity-specific details like teeth outlines (middle row) and stubble (bottom row), while previous methods degenerate into blur or noise. Second, the overall face shape is more accurate, as evidenced by the eyeball in the top row and the cheek silhouette in the middle row. Our synthetic face prior encourages the reconstruction to remain faithful to its learned understanding of faces, e.g., that corneas should bulge outwards, not be flat. These results and the metrics in Tbl. 1 demonstrate that our new model and fine-tuning method outperforms previous work both qualitatively and quantitatively.

4.3 Ablation Study

We conduct extensive ablation studies of the prior model. Table 2 lists metrics after fine-tuning to three inputs at 2K resolution. Please see the supp. PDF for more ablations and the HTML page for visuals.

Dataset Size. We compare prior models without any pre-training (a.i) with models pre-trained on a single subject (a.ii), on 15 subjects (a.iii), and on 1,500 subjects (a.iv). Note that each subject is rendered with 13 expressions and 30 views per expression (as described in Sec. 3.2). Hence, the model trained on 1 subject (a.ii) sees $1 \cdot 13 \cdot 30 = 390$ images in total. Without pre-training (a.i), the reconstruction completely fails. Pre-training on a single subject (a.ii) leads to a noisy face geometry (see the surface normals in the supp. HTML page). The performance improves when more subjects are added (a.iii and a.iv). We conclude that pre-training is necessary and more data improves the reconstruction quality.

Number of Pre-training Steps. We ablate the performance when pre-training for a fewer number of steps. We observe that pre-training for one day only (105K steps, b.i) already achieves very high photo-metric reconstruction quality (PSNR and SSIM). This shorter training duration offers a more accessible alternative while

still delivering high-quality outcomes. While short pre-training already yields good results, longer training is required for the best perceptual quality (LPIPS) (b.ii and b.iii).

Synthetic Data Quality. We investigate the effects of synthetic data quality in terms of appearance and texture (c.i - c.iv) and geometric diversity (c.v - c.vii). We find that the appearance and texture of the prior model have very little impact on the fine-tuning result but a higher geometry diversity achieves the best results. Prior models trained on gray-scale and low-quality renderings (c.i and c.ii) perform similarly as training in diverse environments (c.iii) [Gardner et al. 2017; Hold-Geoffroy et al. 2019]. Similarly, excluding details like makeup (c.iv) does not deteriorate the performance. However, the diversity in terms of geometry is important for the prior model. When removing hair (c.vi) and other accessories like beards, glasses, and clothing (c.v) from the prior model, the reconstruction still yields a valid face but it may contain some artifacts in non-surface regions like hair. This is notable as a drop in PSNR from 26.54 with all accessories (c.vii) to 26.00 without any accessories (c.v) (please see the supp. mat. for a list of the number of accessories). In summary, we find that the prior acts as a geometric regularization. Including all accessories yields a geometrically diverse prior model with the best results. Rendering even more assets is likely to improve the results for accessories like glasses, earrings, and clothing even further.

Including Real Data. We study the synthetic vs. real domain gap by pre-training the prior model on real multi-view images (d.i), a mixed dataset with real and synthetic images (d.ii), and synthetic images alone (d.iii). Priors trained on a real and a mixed dataset perform well but synthetic data alone performs best on all metrics. This behavior might seem non-intuitive. While it is very difficult to precisely determine why synthetic data outperforms real data, we are not the first to find that synthetic data can outperform real data [Sun et al. 2021; Trevithick et al. 2023; Wood et al. 2021, 2022; Yeh et al. 2022]. In our experience, real data capture and processing is imperfect compared with synthetic data. Even under controlled conditions, there may be issues like motion blur and imperfect foreground matting.

4.4 Limitations

While our method can reconstruct 3D faces from even just one view, we notice that the quality degrades at side views in this extremely challenging case (see Fig. 10) due to the lack of observation. Accessories like glasses frames and earrings may not be reconstructed perfectly and for extreme expressions, the mouth interior might not be fully consistent across all viewpoints (see the supp. HTML page). Furthermore, our method assumes that faces are non-occluded in the input images. This can cause the camera estimation during the 3DMM fitting (Sec. 3.3.1) to fail. These limitations could be potentially addressed by leveraging large generative models to hallucinate unobserved regions, which is an interesting future direction to explore. Another direction of future work could explore more efficient backbones to increase pre-training and fine-tuning efficiency, and potentially enable facial animation.

5 CONCLUSION

We present a method for high-fidelity expressive face modeling from as few as three input views captured in the wild. This challenging goal is achieved by leveraging a volumetric face prior and fine-tuning the prior model to sparse observations. Our key insight is that a prior model trained on synthetic data alone can generalize to diverse real-world identities and expressions, bypassing the expensive process of capturing large-scale real-world 3D facial appearances. We experimentally demonstrate that our method can robustly reconstruct expressive faces from sparse or even single-view images with unprecedented fine-scale idiosyncratic details, and achieve superior quality compared to previous state-of-the-art methods for few-shot reconstruction and novel view synthesis.

REFERENCES

- Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5855–5864.
- Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5470–5479.
- Jan Bednarik, Erroll Wood, Vassilis Choutas, Timo Bolkart, Daoye Wang, Chenglei Wu, and Thabo Beeler. 2024. Learning to Stabilize Faces. *Computer Graphics Forum* (2024). <https://doi.org/10.1111/cgf.15038>
- 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.
- Yochai Blau and Tomer Michaeli. 2018. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6228–6237.
- Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. 2018. Optimizing the latent space of generative networks. In *Proceedings of the 35th International Conference on Machine Learning*. 2640–3498.
- Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. 2021. VariTex: Variational Neural Face Textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Marcel C Buehler, 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, Shoubo Yu, et al. 2022. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.
- Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16123–16133.
- Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. 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*. 5799–5809.
- Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14124–14133.
- Yunjeon Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8188–8197.
- Yu Deng, Jialong Yang, Jianfeng Xiang, and Xin Tong. 2022a. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Yu Deng, Jialong Yang, Jianfeng Xiang, and Xin Tong. 2022b. GRAM: Generative Radiance Manifolds for 3D-Aware Image Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Hao-Bin Duan, Miao Wang, Jin-Chuan Shi, Xu-Chuan Chen, and Yan-Pei Cao. 2023. Bakedavatar: Baking neural fields for real-time head avatar synthesis. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–17.

- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.
- Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. 2020. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903* (2020).
- Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gabbaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.* 36, 6, Article 176 (nov 2017), 14 pages. <https://doi.org/10.1145/3130800.3130891>
- Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).
- Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. *Technical Report* (2023).
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6927–6935.
- Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5885–5894.
- Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4401–4410.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.
- Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. 2023. NeRsemble: Multi-View Radiance Field Reconstruction of Human Heads. *ACM Trans. Graph.* 42, 4, Article 161 (jul 2023), 14 pages. <https://doi.org/10.1145/3592455>
- Mariska E. Kret. 2015. Emotional expressions beyond facial muscle actions. A call for studying autonomic signals and their impact on social perception. *Frontiers in Psychology* 6 (2015). <https://doi.org/10.3389/fpsyg.2015.00711>
- Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. 2022. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12871–12881.
- Daniel H Lee and Adam K Anderson. 2017. Reading what the mind thinks from how the eye sees. *Psychological science* 28, 4 (2017), 494–503.
- Gengyan Li, Kripasindhu Sarkar, Abhimitha Meka, Marcel Buehler, Franziska Mueller, Paulo Gotardo, Otmar Hilliges, and Thabo Beeler. 2024. ShellNeRF: Learning a Controllable High-resolution Model of the Eye and Periocular Region. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15041.
- Ruoshi Liu, Rundui Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9298–9309.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August 15, 2018* (2018), 11.
- Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages. <https://doi.org/10.1145/3450626.3459863>
- Armand Comas Massague, Di Qiu, Menglei Chai, Marcel C. Buehler, Amit Raj, Ruiqi Gao, Qiangeng Xu, Mark Matthews, Paulo F. U. Gotardo, Octavia Camps, Sergio Orts, and Thabo Beeler. 2024. MagicMirror: Fast and High-Quality Avatar Generation with a Constrained Search Space. In *Proceedings of the European conference on computer vision (ECCV)*.
- Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8446–8455.
- Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. 2022. KeypointNeRF: Generalizing Image-based Volumetric Avatars using Relative Spatial Encoding of Keypoints. In *European conference on computer vision*.
- 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*. Springer, 405–421.
- Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).
- Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis from Sparse Inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.
- Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. 2022. StyleSDF: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13503–13513.
- Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–21.
- Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, and Stefanos Zafeiriou. 2023. Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8806–8817.
- Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. 2021a. Nerfies: Deformable Neural Radiance Fields. *ICCV* (2021).
- Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. 2021b. HyperNeRF: A Higher-Dimensional Representation for Topologically Varying Neural Radiance Fields. *ACM Trans. Graph.* 40, 6, Article 238 (dec 2021).
- Olivier Pascalis and David J Kelly. 2009. The origins of face processing in humans: Phylogeny and ontogeny. *Perspectives on psychological science* 4, 2 (2009), 200–209.
- Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. *arXiv* (2022).
- Malte Prinzler, Otmar Hilliges, and Justus Thies. 2023. Diner: Depth-aware image-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12449–12459.
- Eduard Ramon, Gil Triginer, Janna Ecur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. 2021. H3D-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5620–5629.
- Pramod Rao, Mallikarjun B R, Gereon Fox, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Ayush Tewari, Christian Theobalt, and Mohamed Elgharib. 2022. VoRF: Volumetric Relightable Faces. *British Machine Vision Conference (BMVC)* (2022).
- Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. 2022. Lohnerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1558–1567.
- Kripasindhu Sarkar, Marcel C Buehler, Gengyan Li, Daoye Wang, Delio Vicini, Jérémy Riviere, Yinda Zhang, Sergio Orts-Escolano, Paulo Gotardo, Thabo Beeler, et al. 2023. LitNeRF: Intrinsic Radiance Decomposition for High-Quality View Synthesis and Relighting of Faces. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.
- Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.
- Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*.
- Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. 2006. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proc. IEEE* 94, 11 (2006), 1948–1962. <https://doi.org/10.1109/JPROC.2006.884093>
- Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. 2020. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems* 33 (2020), 10136–10147.
- Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. 2021. NeLF: Neural Light-transport Field for Portrait View Synthesis and Relighting. In *Eurographics Symposium on Rendering*.
- Feitong Tan, Sean Fanello, Abhimitha Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. 2022. VoLux-GAN: A Generative Model for 3D Face Synthesis with HDRI Relighting. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) (SIGGRAPH '22). Association for Computing Machinery, New York, NY, USA, Article 58, 9 pages. <https://doi.org/10.1145/3528233.3530751>
- Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. 2021. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2846–2855.
- Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. 2023. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 22819–22829.

- Alex Trevithick, Matthew Chan, Michael Stengel, Eric R. Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Mannohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. 2023. Real-Time Radiance Fields for Single-Image Portrait View Synthesis. In *ACM Transactions on Graphics (SIGGRAPH)*.
- Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. 2023. SPARF: Neural Radiance Fields from Sparse and Noisy Poses. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*.
- Vishal Vinod, Tanmay Shah, and Dmitry Lagun. 2024. TEGLO: High Fidelity Canonical Texture Mapping from Single-View Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3585–3595.
- Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. 2021. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260* (2021).
- 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* (Vancouver, BC, Canada) (SIGGRAPH '22). Association for Computing Machinery, New York, NY, USA, Article 55, 9 pages. <https://doi.org/10.1145/3528233.3530753>
- Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. IBRNet: Learning Multi-View Image-Based Rendering. In *CVPR*.
- Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrušaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, et al. 2023. Rodin: A generative model for sculpting 3d digital avatars using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4563–4573.
- Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. 2024. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems* 36 (2024).
- Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. 2021. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3681–3691.
- Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 2022. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*. Springer, 160–177.
- Rundi Wu, Ben Mildenhall, Philipp Hertzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21551–21561.
- Cheng-hsin Wu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, 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, Shouo-I Yu, and Yaser Sheikh. 2022. Multiface: A Dataset for Neural Face Rendering. In *arXiv*. <https://doi.org/10.48550/ARXIV.2207.11243>
- Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*. Springer, 736–753.
- 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.
- Jiawei Yang, Marco Pavone, and Yue Wang. 2023. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.
- Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Transactions on Graphics (TOG)* (2022).
- Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields from One or Few Images. In *CVPR*.
- Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. 2019. Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. 2023. Avatarbooth: High-quality and customizable 3d human avatar generation. *arXiv preprint arXiv:2306.09864* (2023).
- Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. 2022. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. In *SIGGRAPH Asia 2022 Conference Papers*. 1–9.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.
- Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. 2023. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Transactions on Graphics* 43, 1 (2023), 1–16.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.
- Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 21057–21067.
- Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. 2021. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. (2021). *arXiv:2110.09788*
- Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Menghua Wu, Qiu Shen, Ruigang Yang, and Xun Cao. 2023. FaceScape: 3D Facial Dataset and Benchmark for Single-View 3D Face Reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2023).

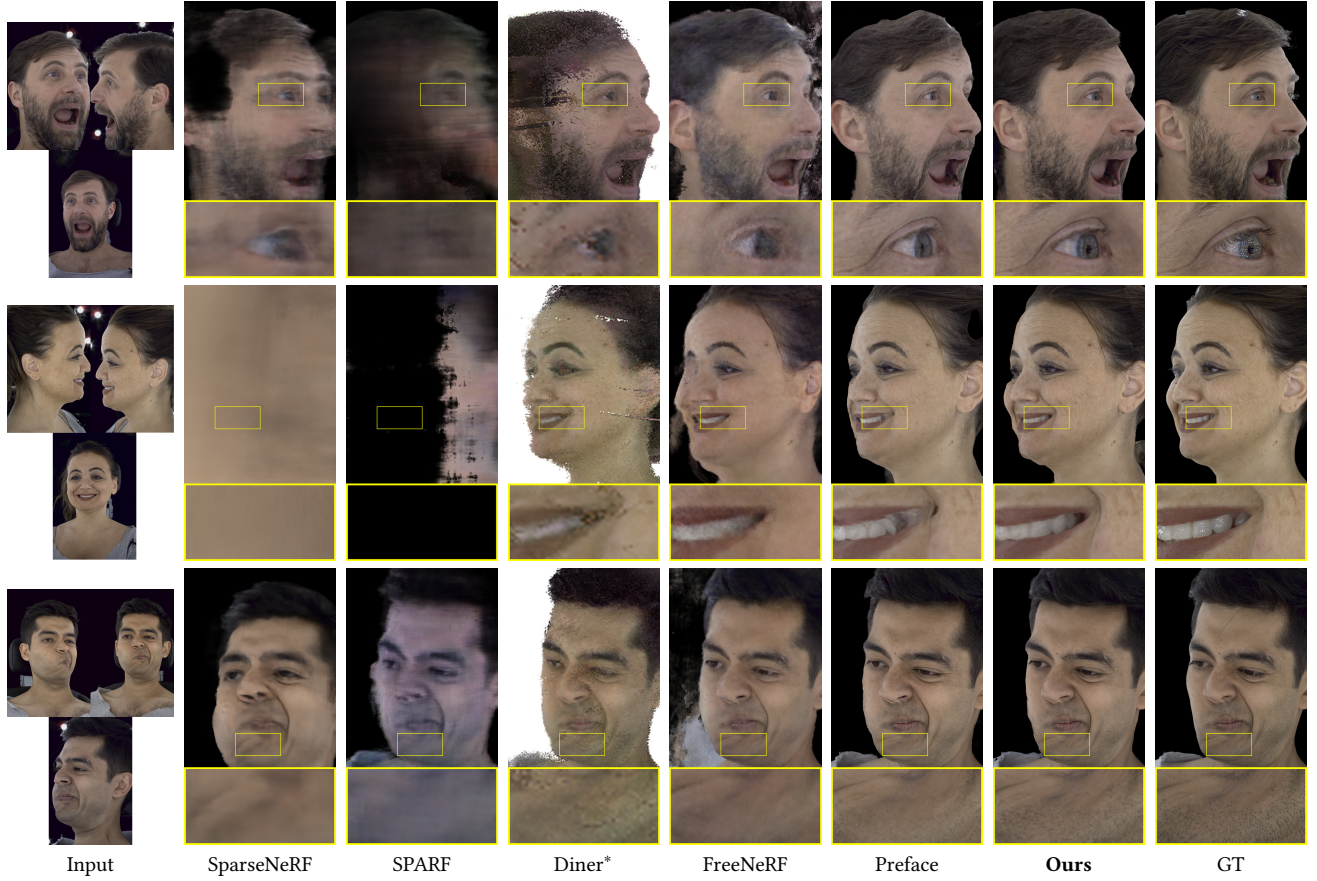


Fig. 7. We compare our method with previous work on the Multiface dataset [Wuu et al. 2022]. For each expression, we show the novel-view results with three input views. The symbol * indicates results at a lower resolution. Both visually and quantitatively, our method significantly outperforms SparseNeRF [Guangcong et al. 2023], SPARF [Truong et al. 2023], Diner [Prinzler et al. 2023], and FreeNeRF [Yang et al. 2023]. Compared to the state-of-the-art Preface [Buehler et al. 2023], we see a clear improvement in areas like eye, chin, and face contours, as shown in the zoom-in. For example, our method better reconstructs the teeth on row 2. We also compare with Preface on in-the-wild captures, please see Fig. 8.

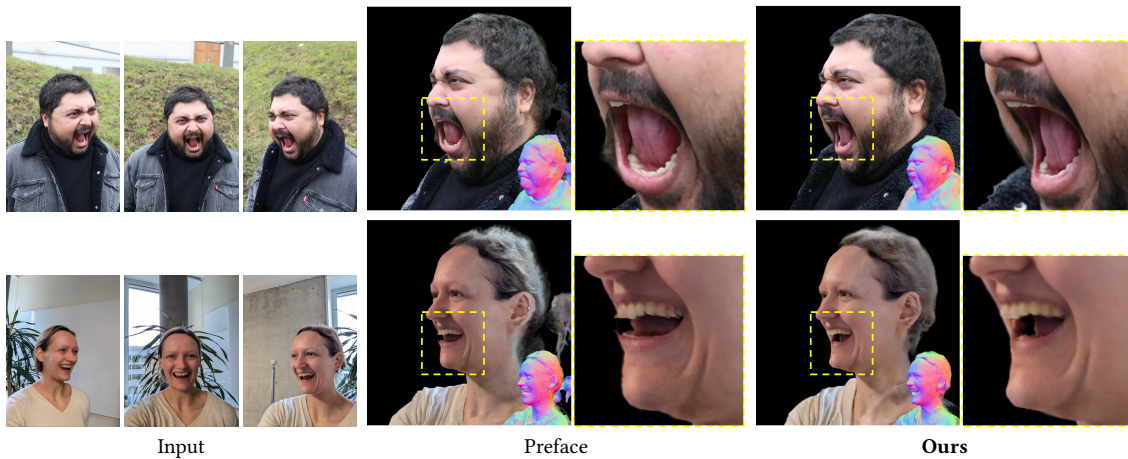


Fig. 8. We highlight a typical failure case for the best-performing related work. Preface struggles [Buehler et al. 2023] with strong facial expressions, in particular in the mouth and chin region. Please see the supplementary HTML page for video results.



Fig. 9. In-the-wild results. Given three in-the-wild images of a subject, our method reconstructs the subject and renders novel views with high resolution. The figure shows the input images (which can be taken sequentially), high-resolution rendering results in novel view, and also the reconstructed normal and depth maps. Our model generalizes to different challenging in-the-wild lighting. These results include indoor, outdoor, and dim scenes. Our method also captures strong idiosyncratic facial expressions such as the tongue-out case in row 3. This is an expression not included in the synthetic training data.

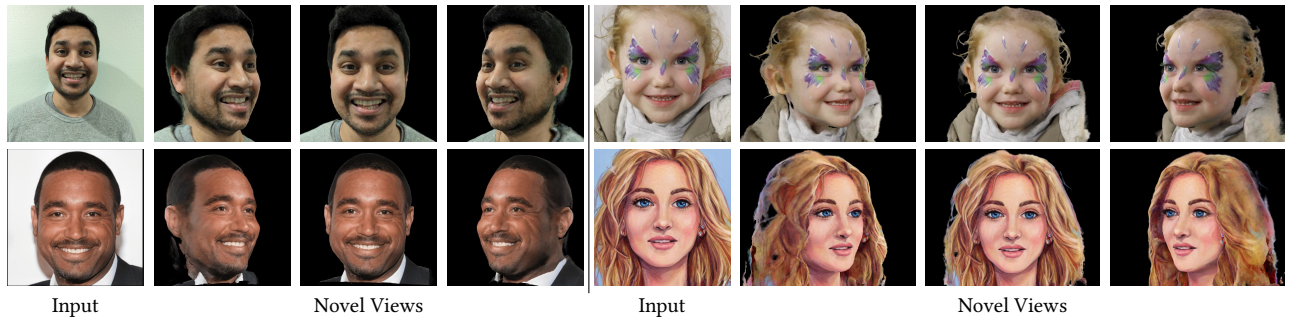


Fig. 10. Single Image Results. We push the limits of our method by reconstructing the face with only a single input image. The top left image is a smartphone image, the bottom left an Image from Wang et al. (2023) [Wang et al. 2023], and the right two examples are stylized images from [Trevithick et al. 2023].

A SUPPLEMENTARY

This supplementary document provides more details about the synthetic dataset (Sec. A.1), the models and method (Sec. A.2), experiments (Sec. A.4), and supplementary results (Sec. A.5).

A.1 Dataset Details

Our synthetic dataset is generated in a similar fashion to previous work [Wood et al. 2021]. We use Blender to both generate and render realistic and diverse 3D scenes containing a face. We sample faces from a large dataset of 50,000 3D scans that have been registered with a common template mesh [Bednarik et al. 2024]. Then, we make each face mesh look realistic by applying a physically-based skin material from a high-resolution skin texture collection. Next, we “dress up” our face by procedurally attaching traditional CG assets¹. Our total asset library contains the following items: 23 upper-body garments (jackets, sweaters, suits, uniforms, coats, t-shirts, scarfs),

8 types of eye-glasses, 157 eyebrows, 6 types of makeup, 2 sets of eyelashes (with and without mascara), 1,679 head textures, 2 headwear items, 125 strand-based hair / beard styles, and 6 eye textures. Each asset is authored as a rig using Blender Geometry Nodes, and each class of item executes class-specific rig logic to robustly attach itself to arbitrary 3D face meshes, regardless of identity or expression. For example, upper-body clothing items non-rigidly deform themselves to fit around the neck of a target face, but eyeglasses are posed using inverse kinematics to rest on the nose-bridge and ears.

Finally, we render the scenes with Cycles: a physically-based ray-tracing renderer. We render all faces in the same environment: a uniform well-lit environment. Each face is rendered from 30 distinct viewpoints, generated by placing the camera at random points around the head and pointing it at the face. We choose random camera positions by sampling spherical coordinates: azimuthal angle, elevation angle, and radius. We avoid overly similar viewpoints by discarding those with viewing directions closer than 25 degrees to a previous one and re-sampling. We found it helpful to sample random viewpoints for each subject and expression rather than sampling the same viewpoints across the entire dataset. The latter led to more floaters when training the prior model.

For the ablation with diverse environments, we sample random environment maps from the Laval indoor dataset [Gardner et al. 2017; Hold-Geoffroy et al. 2019].

A.2 Method Details

A.2.1 Prior Model Details.

Architecture. The prior model architecture largely follows Preface [Buehler et al. 2023]. The prior model has a *proposal* MLP predicting density and normals, and a *NeRF* MLP predicting density and color [Barron et al. 2022]. The proposal MLP has depth 4 and layers width $(256 + |\beta| + |\psi| + |\mathbf{w}|) \times 256$ parameters, where the $\beta \in \mathbb{R}^{48}$ are the 3DMM identity coefficients, $\psi \in \mathbb{R}^{157}$ are the 3DMM expression coefficients, and $\mathbf{w} \in \mathbb{R}^{64}$ is the optimizable latent identity code. The NeRF MLP trunk has depth 8 and layers width $(1024 + |\beta| + |\psi| + |\mathbf{w}|) \times 1024$. The NeRF MLP trunk features are projected to 3 dimensions and normalized to predict negative normals. The color is predicted from the NeRF MLP trunk features after a bottleneck with width 256 and a single view-conditioned layer with width 128. We optimize one \mathbf{w} code per identity, which yields a codebook of size 1500×64 . The total parameter count is 32 Mio. The weights are initialized with He Uniform Variance Scaling [He et al. 2015].

Training. We train the prior model on images of 1,500 synthetic identities, rendering each with 13 expressions and 30 views. In total, we train the full prior model for 1 Mio. steps on 64 TPUs with a batch size of 131,072 rays per step ($256 \text{ identities} \times 8 \text{ view} \times 64 \text{ pixels}$), which takes about 10 days. However, we observe that the model already reaches near convergence after 105,000 steps. In particular, fine-tuning that uses the model trained for 105,000 steps achieves very similar photo-metric quality as fine-tuning the model trained for 1 Mio. steps, please see our ablation study in Sec. 4.3 and the supp. HTML page for visuals.

¹<https://www.3dscanstore.com/>

We train our model with 128 samples for the proposal and 128 samples for the NeRF MLP. The proposal MLP is sampled twice and the NeRF MLP is sampled once. Both the proposal and NeRF MLP use the same positional encoding for the inputs. The xyz inputs use twelve levels; the view direction four levels and appends the view direction without positional encoding.

Losses. The terms in the loss function $\mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{prop}} \mathcal{L}_{\text{prop}}$ are a combination of the mean absolute photo-metric error of the ground truth color \mathbf{c} versus the predicted color $\hat{\mathbf{c}}$: $\mathcal{L}_{\text{recon}} = \|\mathbf{c} - \hat{\mathbf{c}}\|_1$, and the MipNeRF360 proposal loss $\mathcal{L}_{\text{prop}} = \sum_i \frac{1}{w_i} \max(0, w_i - \text{bound}(\hat{\mathbf{t}}, \hat{\mathbf{w}}, T_i))^2$. Please see Eq. 13 in the original paper for more details [Barron et al. 2022].

The optimization employs Adam [Kingma and Ba 2015] with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate decays exponentially from 0.002 to 0.00002. We clip gradients with norms larger than 0.001.

We also experimented with the supervision of accumulation and depth but didn't find an improvement.

A.3 Inference Details

3DMM Fitting. We fit the 3DMM [Blaiz and Vetter 1999] from 599 2D probabilistic landmarks, where each landmark is the projection of one vertex of the 3DMM mesh with an uncertainty σ . The landmark fitting follows [Wood et al. 2022] and minimizes the energy function

$$E(\Phi; L) = E_{\text{landmarks}} + E_{\text{identity}} + E_{\text{expression}} + E_{\text{joints}} + E_{\text{intersect}},$$

where L denotes the 599 probabilistic landmarks and Φ all the optimized 3DMM parameters including identity, expressions, joint rotations, and global translation, and intrinsic and extrinsic camera parameters, if unknown. The landmark term

$$E_{\text{landmarks}} = \sum_{j,k} \frac{\|\mathbf{x}_{jk} - \boldsymbol{\mu}_{jk}\|^2}{2\sigma_{jk}^2}$$

minimizes the distance between the projected landmarks from the 3DMM \mathbf{x}_{jk} and the estimated probabilistic landmarks with mean $\boldsymbol{\mu}_{jk}$ and variance σ_{jk}^2 for all available camera views C . The identity $E_{\text{identity}} = -\log(p(\boldsymbol{\beta}))$ is a regularizer that encourages plausible face shapes. It is computed as the negative log-likelihood given a fitted Gaussian Mixture Model of 3D head scans [Wood et al. 2021]. The expression and joints terms $E_{\text{expression}} = \|\boldsymbol{\psi}\|^2$ and $E_{\text{joints}} = \|\mathbf{J}_i\|^2$ are regularization terms to encourage small values in the expression code $\boldsymbol{\psi}$ and joint rotations \mathbf{J}_i for the neck and the two eyeball joints. Note that the regularization does not apply to the global (head) joint. $E_{\text{intersect}}$ discourages intersecting vertices. Please see [Wood et al. 2022] for more details.

Warm-up by Inversion. We sample random patches of size 32×32 . Our batch size is 4096 rays so we use 4 patches in each batch. We run 1,500 inversion steps, which takes about 10 minutes on four TPUs.

Optimization. The model fitting uses the same sampling strategy as prior model training (2x proposal and 1x NeRF sampling) and employs the Adam optimizer with the same parameters. We sample 4096 random rays across all available views in each step. We optimize for 50,000 steps, which takes about 3.5 hours.

Losses. The loss terms follow related works. In the following, the variable w corresponds to the NeRF sample weight defined in Eq. 5 of the original NeRF paper [Mildenhall et al. 2020]. The normal consistency loss is defined as $\mathcal{L}_{\text{normal}} = \sum_i w_i \cdot (1 - \mathbf{n}^\top \hat{\mathbf{n}})$, where \mathbf{n} are the analytical normals and $\hat{\mathbf{n}}$ are the predicted normals. The regularization of the view direction weights is defined as $\mathcal{L}_d = \|\theta_v\|^2$, where θ_d are the model parameters that process the positionally encoded view directions. The distortion loss is defined as $\mathcal{L}_{\text{dist}} = \sum_{i,j} w_i w_j \left\| \frac{s_i + s_{j+1}}{2} - \frac{s_j + s_{j+1}}{2} \right\| + \frac{1}{3} \sum_i w_i^2 (s_{i+1} - s_i)$, where s is the normalized ray distance. For details about the LPIPS loss $\mathcal{L}_{\text{LPIPS}}$, please refer to Eq. 1 in [Zhang et al. 2018]. We find that sampling individual random rays during model fitting yields better results than sampling patches. Hence, we only employ the perceptual loss $\mathcal{L}_{\text{LPIPS}}$ during warm-up but not during fine-tuning.

Rendering Time. We render on 4 TPUs, where rendering takes about 20.5 seconds per 1024×1024 frame.

A.4 Experimental Details

A.4.1 Multiface Dataset. We quantitatively compare and ablate on a subset of the Multiface dataset [Wuu et al. 2022]. Metrics are computed on three expressions from three identities—nine scenes in total. For each scene, we select three views for training: One frontal and two side views. For evaluation, we compute metrics on all holdout views where the cameras are not located on the back of the head. Please see Fig. 11 for a visualization.

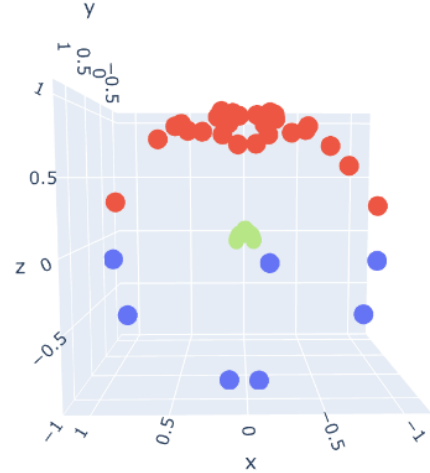


Fig. 11. Multiface cameras used for evaluation are highlighted in red. Blue cameras are discarded. The face (green) is looking along the z axis.

A.4.2 Preface Dataset. Our ablation study that trains on real data uses the Preface dataset [Buehler et al. 2023]. It contains multi-view captures of 1,500 identities. Each identity is captured for 13 facial expressions and from 12 views. Please see [Buehler et al. 2023] for more details and a breakdown of demographics.

# Views	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
1	23.79	0.7323	0.3139
2	24.29	0.7409	0.3184
3	26.54	0.7750	0.3144

Table 3. Number of Input Views. Our method can produce good-quality frontal views from a single frontal image. However, the quality suffers from side views, where the input image doesn’t provide any information. Please see the supplementary HTML page for visuals.

A.5 Supplementary Results

A.5.1 In-the-wild Results. We provide extensive results for high-resolution in-the-wild captures on the supplementary HTML page and video.

Table 4. Additional ablations. Training a prior model with a background leads to more floating artifacts during fine-tuning.

Variant	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
Prior with background	24.51	0.7313	0.3244
Prior without background	26.54	0.7750	0.3144

A.5.2 Supplementary Ablations and Comparisons. We extend our ablations from the main paper with metrics computed on different variants of our prior model in Tables 3 and 4. The metrics are computed on the Multiface dataset [Wuu et al. 2022], as described in the main paper.

We ablate fine-tuning results when a different number of views are available in Tbl. 3. We find that our method produces pleasing front-view faces even for a single input view. However, the quality quickly degrades for side views, where the input views do not provide any signal. Table 4 ablates a prior model trained without background removal. Keeping the background (i) yields more floaters during model fitting.

For visuals and additional results for single image inputs, please see the main paper and the supplementary HTML page.

A.6 Ethics

The content of this paper follows the SIGGRAPH Asia policies for data privacy. In particular, all in-the-wild subjects have signed an agreement to be captured and reconstructed for research purposes. This approach inhibits the same risks and dangers as other face reconstruction methods.