



Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis

Conference Paper**Author(s):**

Bühler, Marcel  Sarkar, Kripasindhu; Shah, Tanmay; Li, Gengyan; Wang, Daoye; Helminger, Leonhard; Orts-Escolano, Sergio; Lagun, Dmitry; Hilliges, Otmar  Beeler, Thabo; Meka, Abhimitra

Publication date:

2023

Permanent link:

<https://doi.org/10.3929/ethz-b-000648001>

Rights / license:

In Copyright - Non-Commercial Use Permitted

Originally published in:

<https://doi.org/10.1109/ICCV51070.2023.00315>

Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis

Marcel C. Bühler^{1,2} Kripasindhu Sarkar² Tanmay Shah² Gengyan Li^{1,2} Daoye Wang²
 Leonhard Helming² Sergio Orts-Escolano² Dmitry Lagun²
 Otmar Hilliges¹ Thabo Beeler² Abhimitra Meka²
¹ETH Zurich ²Google



Figure 1. We propose a method for synthesising novel views of faces at ultra high-resolution from very sparse inputs. This figure shows novel view renderings at **4K resolution** reconstructed from **only three views** of the target identity.

Abstract

NeRFs have enabled highly realistic synthesis of human faces including complex appearance and reflectance effects of hair and skin. These methods typically require a large number of multi-view input images, making the process hardware intensive and cumbersome, limiting applicability to unconstrained settings. We propose a novel volumetric human face prior that enables the synthesis of ultra high-resolution novel views of subjects that are not part of the prior’s training distribution. This prior model consists of an identity-conditioned NeRF, trained on a dataset of low-resolution multi-view images of diverse humans with known camera calibration. A simple sparse landmark-based 3D alignment of the training dataset allows our model to learn a smooth latent space of geometry and appearance despite a limited number of training identities. A high-quality volumetric representation of a novel subject can be obtained by model fitting to 2 or 3 camera views of arbitrary resolution. Importantly, our method requires as few as two views of casually captured images as input at inference time.

1. Introduction

Reconstruction and novel view synthesis of faces are challenging problems in 3D computer vision. Achieving high-quality photorealistic synthesis is difficult due to the underlying complex geometry and light transport effects exhibited by organic surfaces. Traditional techniques use explicit geometry and appearance representations for modeling individual face parts such as hair [14], skin [17], eyes [4], teeth [59] and lips [16]. Such methods often require specialised expertise and hardware and limit the applications to professional use cases.

Recent advances in volumetric modelling [3, 26, 31, 48] have enabled learned, photorealistic view synthesis of both general scenes and specific object categories such as faces from 2D images alone. Such approaches are particularly well-suited to model challenging effects such as hair strands and skin reflectance. The higher dimensionality of the volumetric reconstruction problem is inherently more ambiguous than surface-based methods. Thus, initial developments in neural volumetric rendering methods [3, 31] relied on an order-of-magnitude higher number of input images (> 100)

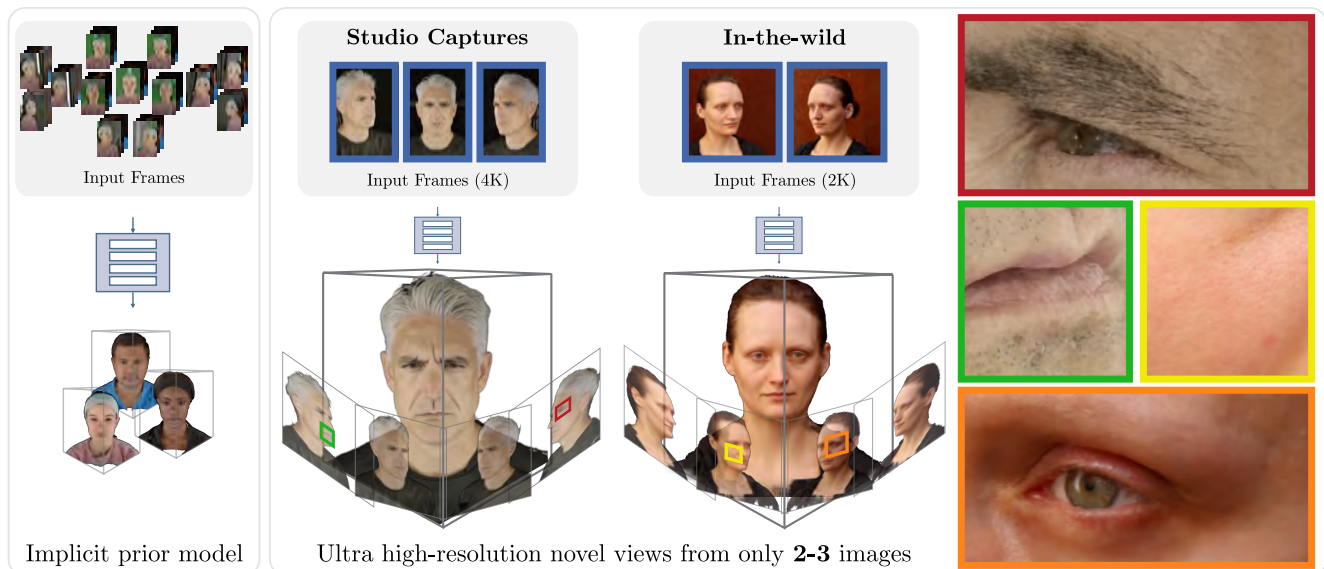


Figure 2. Our key contribution is a prior face model (left), learned from a multiview dataset of faces captured in a controlled setting. The prior model is resolution independent and can be fine-tuned to synthesise novel views at high resolution given as few as two images from a target identity captured in the studio (middle left) or in-the-wild (middle right).

to make the solution tractable. Such a large image acquisition cost limits application to wider casual consumer use cases. Hence, few-shot volumetric reconstruction, of both general scenes and specific object categories such as human faces, remains a prized open problem.

This problem of the inherent ambiguity of volumetric neural reconstruction from few images has generally been approached in 3 ways: i) Regularisation: using natural statistics to constrain the density field better such as low entropy [3, 46] along camera rays, 3D spatial smoothness [35] and deep surfaces [66] to avoid degenerate solutions such as floating artifacts; ii) initialisation: meta-learned initialisation [53] of the underlying representation (network weights) to aid faster and more accurate convergence during optimisation; iii) data-driven subspace priors: using large and diverse datasets to learn generative [7, 9, 10, 12, 13, 18, 68] or reconstructive [6, 44, 46, 57] priors of the scene volume.

For human faces, large in-the-wild datasets [21, 22, 25] have proved to be particularly attractive in learning a smooth, diverse, and differentiable subspace that allow for few-shot reconstruction of novel subjects by performing inversion and finetuning of the model on a small set of images of the target identity [47]. But such general datasets and generative models also suffer from disadvantages: i) The sharp distribution of frontal head poses in these datasets prevents generalisation to more extreme camera views, and ii) the computational challenge of training a 3D volume on such large datasets results in very limited output resolutions.

In this paper, we propose a novel volumetric prior for faces that is learned from a multi-view dataset of diverse



Figure 3. Naively training on two images leads to overfitting and the model fails to synthesise novel views. With the proposed prior, the model can render view-consistent novel views.

human faces. Our model consists of a neural radiance field (NeRF) conditioned on learnt per-identity embeddings trained to generate 3D consistent views from the dataset. We perform a pre-processing step that aligns the geometry of the captured subjects [46]. This geometric alignment of the training identities allows our prior model to learn a continuous latent space using only image reconstruction losses. At test time, we perform model inversion to compute the embedding for a novel target identity from the given small set of views of arbitrary high resolution. In an out-of-model finetuning step, the resulting embedding and model are further trained with the given images. This results in NeRF model of the target subject that can synthesise high-quality images. Without our prior, the model cannot estimate a 3D consistent volume and overfits to the sparse training views (Fig. 3).

While we present a novel data-driven subspace prior, we also extensively evaluate the role of regularisation and initialisation in achieving plausible 3D face volumes from few images by comparing with relevant state-of-the-art techniques and performing design ablations of our method.

In summary, we contribute:

- A prior model for faces that can be finetuned to generate a high-quality volumetric 3D representation of a target identity from two or more views.
- Ultra high-resolution 3D consistent view-synthesis (demonstrated up to 4k resolution).
- Generalisation to in-the-wild indoor and outdoor captures, including challenging lighting conditions.

2. Related works

Volumetric reconstruction techniques [3, 24, 32, 40, 41] achieve a high-level of photorealism. However, they provide a wider space of solutions than surface based representations [29, 39], and hence often perform very poorly in the absence of sufficient constraints [30, 35, 42, 54, 62, 63]. To mitigate this, related works employ additional regularisation [20, 30, 35, 46, 54, 62], perform sophisticated initialisation [23, 44, 53, 56], and leverage data-driven priors [9, 11, 18, 20, 30, 37, 42, 44, 46, 52, 54, 58, 63, 65].

Regularisation A common solution to novel view synthesis from sparse views is employing regularisation and consistency losses for novel views.

RegNeRF [35] proposes a smoothness regulariser on the expected depth and a patch-based appearance regularisation from a pretrained normalising flow. A concurrent work, FreeNeRF [62], observes that NeRFs tend to overfit early in training because of the high frequencies in the positional encoding. They propose a training schedule where the training starts with the positional encodings masked to the low frequencies only and continuously fade in higher frequencies during the course of training. These methods have shown promising results for in-the-wild scenes but struggle to output high-quality results for human faces 8.

It is also possible to leverage priors from large pretrained models. DietNeRF[20] follows a strategy of constraining high-level semantic features of novel view images to map to the same scene object in the “CLIP” [43] space. These methods require generating image patches per mini-batch rather than individual pixels. This is compute and memory intensive and reduces the effective batch size and resolution at which the models can be trained, limiting the overall quality.

Initialisation Recent papers explore the effect of initialisation [23, 44, 53, 56]. Metalearning [15, 33, 51, 64] initial model parameters from a large collection of images [53] has shown promising results for faster convergence. However, the inner update loop in metalearning becomes very expensive for large neural networks. This limits its applicability in high-resolution settings.

Data-driven Priors Recent works propose generative neural fields models in 3D [7, 9, 12, 18, 36, 45, 46, 50, 52, 60, 68]. These models typically map a random latent vector to a radiance field. At inference time, the model can generate novel views by inverting a target image to the latent space [1].

GRAF and PiGAN [7, 50] are the first technique to learn a 3D volumetric generative model trained with an adversarial loss on in-the-wild datasets. Since neural radiance fields are computationally expensive, training them in an adversarial setting requires an efficient representation. EG3D [9] proposes a tri-plane representation, which enables training lightweight neural radiance field as a 3D GANs, resulting in state-of-the-art synthesis results.

Due to memory limitations, such generative models can be trained only at limited resolutions. They commonly rely on an additional 2D super-resolution module to generate more details [7, 9, 18, 52], which results in the loss of 3D consistency.

Recent works render 3D consistent views by avoiding a 2D super-resolution module [6, 57]. MoRF [57] learns a conditional NeRF [32] for human heads from multiview images captured using a polarisation based studio setup that helps to learn separate diffuse and specular image components. Their dataset consists of 15 real identities and is supplemented with synthetic renderings to generate more views. Their method is limited to generating results in the studio setting and does not generalise to in-the-wild scenes. Cao et al. 2022 [6] train a universal avatar prior that can be finetuned to a target subject with a short mobile phone capture of RGB and depth. Their underlying representation follows Lombardi et al. [27].

A popular option for novel view synthesis from sparse inputs is formulating the task as an auto-encoder and perform image-based rendering. This family of methods [11, 30, 58, 63] follow a feedforward approach of generalisation to novel scenes by training a convolutional encoder that maps input images to pixel aligned features that condition a volumetric representation of the scene.

Multiple works extend this approach with additional priors including keypoints [30], depth maps [19, 42, 61], or correspondences [54]. KeypointNeRF [30] employs an adapted positional encoding strategy based on 3D keypoints. DINER [42] includes depth maps estimated from pretrained models to bootstrap the learning of density field

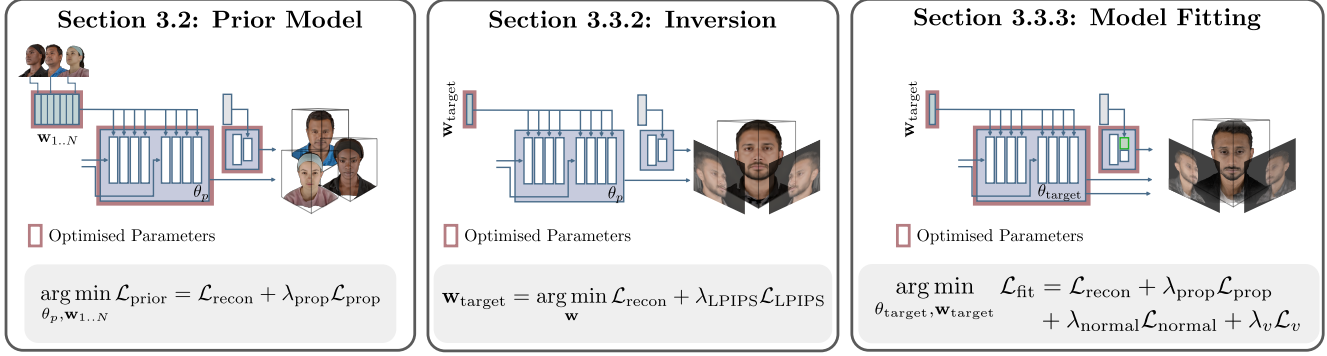


Figure 4. Overview. We train an implicit prior model on low-resolution multi-view images (left). At test time, we fit the prior model to as few as two images of a target identity. A naïve optimisation without inversion or regularisation leads to strong view-dependent colour distortions and fuzzy surface structures, see Sec. 6.4 and Fig. 11. To solve this, we first find a good initialisation through inversion (middle) and then finetune all model parameters under additional constraints for geometry $\mathcal{L}_{\text{normal}}$ and appearance \mathcal{L}_v (right).

and sample the volume more efficiently around the expected depth value. Employing our face prior outperforms these methods (see Tbl. 1, Fig. 8 and 9).

3. Method

We propose a prior model for faces that can be finetuned to very sparse views. The finetuned model can generate ultra-high resolution novel view synthesis with intricate details like individual hair strands, eyelashes, and skin pores (Fig. 1). In this section, we first introduce neural radiance fields [32] in Sec. 3.1 and our prior model in Sec. 3.2. We then outline our reconstruction pipeline in Sec. 3.3.

3.1. Background

A NeRF [31] represents a scene as a volumetric function $f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ which maps 3D locations \mathbf{x} to a radiance \mathbf{c} and a density σ , which is modelled using a multi-layer perceptron (MLP). The radiance is additionally conditioned on the view direction \mathbf{d} to support view dependent effects such as specularities. In order to more effectively represent and learn high frequency effects, each location is positionally encoded before being passed to the MLP.

Given a NeRF, a pixel can be rendered by integrating along its corresponding camera ray in order to obtain the radiance or colour value $\hat{\mathbf{c}} = \mathbf{F}(\mathbf{r})$. Assuming a predetermined near and far camera plane t_n and t_f , the integrated radiance of the camera ray can be computed using the following equation:

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

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

In practice, this is estimated using raymarching. The original NeRF implementation approximated the ray into a discrete number of sample points, and estimated the alpha

value of each sample by multiplying its density with the distance to the next sample. They further improve quality using a coarse-to-fine rendering method, by first distributing samples uniformly between the near and far planes, and then importance sampling the quadrature weights.

Mip-NeRF [2] solves the classic anti-aliasing problem resulting from discrete sampling in a continuous space. This is achieved by sampling conical volumes along the ray. MipNeRF360 [3] also introduced an efficient pre-rendering step; a uniformly sampled coarse rendering pass by a proposal network, which predicts the sampling weights instead of the density and colour values using a lightweight MLP. This is followed by an importance-sampled NeRF rendering step. We incorporate both of these ideas in our model.

3.2. Face Prior Model

Our prior model is a conditional neural radiance field F_θ that is trained as an auto-decoder [5, 46]. Given a ray \mathbf{r} and a latent code \mathbf{w} , F_θ predicts a colour $\hat{\mathbf{c}} = \mathbf{F}_\theta(\mathbf{r}, \mathbf{w})$ with volumetric rendering [32].

The architecture of the prior model is based on Mip-NeRF360 [3] and consists of two MLPs. Unlike Mip-NeRF360, the MLPs are conditioned on a latent code $\mathbf{w}_{\text{identity}}$, representing the identity.

The first MLP—the *proposal* network—predicts density only. The second MLP—the *NeRF* MLP—predicts both density and colour. Both MLPs take an encoded point $\tilde{\gamma}_{\mathbf{x}}(\mathbf{x})$ and a latent code \mathbf{w} as input, where $\tilde{\gamma}_{\mathbf{x}}(\cdot)$ denotes a function for integrated positional encodings [2]. The NeRF MLP further takes the positionally encoded view direction $\gamma_{\mathbf{v}}(\mathbf{d})$ as input (without integration for the positional encoding).

Fig. 5 gives an overview of the backbone NeRF MLP of our prior model. The latent code is concatenated at each layer. Unlike state-of-the-art generative models [8, 18, 46], our model also conditions on the view direction \mathbf{d} .

For training, we sample random rays \mathbf{r} and render the output colour $\hat{\mathbf{c}}$ as described in Sec. 3.1. Given N training

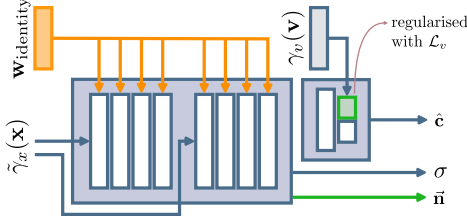


Figure 5. Prior Model Architecture. Our prior model extends the Mip-NeRF360 [3] architecture with a conditioning input at each layer of the trunk MLP. Unlike SOTA generative NeRF models [9, 18, 46], our model conditions both on a latent code *and* a view direction, which enables view-dependent effects. During model fitting to very few images, we prevent overfitting by regularising the view direction weights. See Fig. 11 for an example.

subjects, we optimise over both the network parameters θ and the latent codes $\mathbf{w}_{1..N}$. Our objective function is

$$\arg \min_{\theta, \mathbf{w}_{1..N}} \mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{prop}} \mathcal{L}_{\text{prop}}, \quad (3)$$

with $\lambda_{\text{prop}} = 1$. We describe the loss terms $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{prop}}$ for a single ray. The final loss is computed as the expectation over all rays in the training batch.

The objective function has a data term comparing the predicted colour with the ground truth $\mathcal{L}_{\text{recon}} = \|\mathbf{F}_{\theta}(\mathbf{r}, \mathbf{w}) - \mathbf{c}\|_1$, as well as a weight distribution matching loss term between the NeRF MLP and the proposal MLP $\mathcal{L}_{\text{prop}}$. The latter is the same as in Mip-NeRF360 [3]. We refrain from regularising the latent space, and we disable the distortion loss. As our scene is not unbounded, we also disable the 360-parameterisation or space-warping of Mip-NeRF360.

We train the prior model for 1 Mio. steps on multi-view images of resolution 512×768 . Please refer to Sec. 4 for details about the training set.

3.3. Volumetric Reconstruction Pipeline

Figure 4 illustrates the reconstruction pipeline, which comprises three steps: 1) Preprocessing and head alignment, 2) inversion, and 3) model fitting. This section describes each step in detail.

3.3.1 Preprocessing

We estimate camera parameters and align the heads to a predefined canonical pose during the data preprocessing stage. For the studio setting, we calibrate the cameras and estimate 3D keypoints by triangulating detected 2D keypoints; for in-the-wild captures, we use Mediapipe [28] to estimate the camera positions and 3D keypoints. We align and compute a similarity transform to a predefined set of five 3D keypoints (outer eye corners, nose, mouth centre, and the chin) in a canonical pose. Please see the supp. mat. for details.

3.3.2 Inversion

The reconstruction results depend on a good initialisation of the face geometry (see Tbl. 2). We solve an optimisation problem to find a latent code that produces a good starting point [1].

Given K views of a target identity, we optimise with respect to a new latent code while keeping the network weights frozen. Let P be a random patch sampled from one of the K images of the target identity and $\hat{P}_{\mathbf{w}}$ be a patch rendered by our prior model when conditioning on the latent code \mathbf{w} . The latent code of the target identity $\mathbf{w}_{\text{target}}$ is recovered by minimising the following objective function:

$$\mathbf{w}_{\text{target}} = \arg \min_{\mathbf{w}} \mathcal{L}_{\text{recon}} + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}, \quad (4)$$

where $\mathcal{L}_{\text{recon}} = \frac{1}{|P|} \|\hat{P}_{\mathbf{w}} - P\|$ is the same loss as in Eq. 3, but computed over an image patch, and $\mathcal{L}_{\text{LPIPS}}(\hat{P}_{\mathbf{w}}, P)$ is a perceptual loss[67] with $\lambda_{\text{LPIPS}} = 0.2$. We optimise at the same resolution as the prior model after removing the background [38].

3.3.3 Model Fitting

The goal of model fitting is to adapt the weights of the prior model for generating novel views of a target identity at high resolutions. We do this by finetuning the weights of the prior model to a target identity from sparse views.

Please note that the prior model is trained on *low resolution* and is optimised to reconstruct a *large set of identities* from *many views* for each identity, see Sec. 5. After model fitting, the model should generate *high-resolution novel views* with intricate details like individual hair strands for a *single* target identity given as few as *two* views.

Training a NeRF model on sparse views leads to major artifacts because of a distorted geometry [34] and overfitting to high frequencies [62]. We find that correctly initialising the weights of the model avoids floater artifacts and leads to high-quality novel view synthesis. We initialise the model weights with the pretrained prior model and use the latent code $\mathbf{w}_{\text{target}}$ obtained through inversion (Sec. 3.3.2). Fig. 11 shows that naïvely optimising without any further constraints leads to overfitting to the view direction (first column). Regularising the weights of the view branch causes fuzzy surface structures (second column), which can be mitigated using a normal consistency loss [55] (third column). We initialise the model with the weights of the prior and optimise it given the objective function

$$\arg \min_{\theta_{\text{target}}, \mathbf{w}_{\text{target}}} \mathcal{L}_{\text{fit}} = \mathcal{L}_{\text{recon}} + \lambda_{\text{prop}} \mathcal{L}_{\text{prop}} + \lambda_{\text{normal}} \mathcal{L}_{\text{normal}} + \lambda_v \mathcal{L}_v, \quad (5)$$



Figure 6. Exemplar images of our captured dataset. Our dataset contains 1450 different subjects (*bottom row*) captured under 13 different cameras on the frontal hemisphere (*top rows*).

where the loss terms $\mathcal{L}_{\text{recon}}$, $\mathcal{L}_{\text{prop}}$ and the hyperparameter λ_{prop} are the same as in Eq. 3. The regulariser for the normals $\mathcal{L}_{\text{normal}}$ is the same as in RefNeRF [55]. We regularise the weights of the view branch with $\mathcal{L}_v = \|\theta_v\|^2$, where the parameters θ_v correspond to weights of the connections between the encoded view direction and the output, see the highlighted box in Fig. 5. We set $\lambda_{\text{normal}} = 0.001$ and $\lambda_v = 0.0001$ and optimise until convergence.

Since our model generates faces that are aligned to a canonical pose and location (Sec. 5), the rendering volume can be bounded by a rectangular box. We set the density outside this box to zero for the final rendering.

4. Dataset

We capture a novel high-quality multi-view dataset of diverse human faces from 1450 identities with a neutral facial expression under uniform illumination, see Fig. 6. 13 camera views are distributed uniformly across the frontal hemisphere. Camera calibration is performed prior to every take to obtain accurate camera poses. We hold out 15 identities for evaluation and train on the rest. The camera images are of 4096×6144 resolution. We made a concerted effort for a diverse representation of different demographic categories in our dataset, but acknowledge the logistical challenges in achieving an entirely equitable distribution. We provide more details of the demographic breakdown of the dataset in the supplementary document.

To assess the out-of-distribution performance of our method we show results on the publicly available Facescape multi-view dataset [69]. We also acquire a handful of in-the-wild captures of subjects using a mobile camera to qualitatively demonstrate the generalisation capability of our method further.

5. Experiments

Preprocessing We perform an offline head alignment to a canonical location and pose. This step is crucial to learn a meaningful prior over human faces. For each subject, we es-

| Method | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow |
|----------------------|-----------------|-----------------|--------------------|
| FreeNeRF [62] | 15.02 | 0.6795 | 0.3093 |
| EG3D-based prior [9] | 19.70 | 0.7588 | 0.2897 |
| Learnit [53] | 20.04 | 0.7716 | 0.3299 |
| RegNeRF [34] | 20.40 | 0.7432 | 0.2858 |
| KeypointNeRF [30] | 22.79 | 0.7878 | 0.2713 |
| Ours | 25.69 | 0.8039 | 0.1905 |

Table 1. Comparison with related works at 1K resolution on *two* views of our studio dataset. The metrics are computed as the average over six views of three holdout subjects. Our method outperforms the related works by a clear margin. For a visual comparison, please refer to Fig. 8. The supp. mat. contains metrics and visuals for more input views.

timate five 3D keypoints for the eyes, nose, mouth, and chin and align the head to a canonical location and orientation. The canonical location is defined as the median location of the five keypoints across the first 260 identities of our training set. For an illustration and more details, please see the supplementary document.

Prior Model Training We train the prior model with our pre-processed dataset containing 1450 identities and 13 camera views. To make our training computationally tractable, we train versions of our prior model at a lower resolution. We train two versions of our model, at 256×384 and 512×768 image resolution. The lower resolution model is trained only for the purpose of quantitative evaluation against other SOTA methods, to ensure fair comparison against other methods that cannot be trained at a higher resolution due to compute and memory limitations. We provide details about our training hardware and hyperparameters in the supplementary document.

Comparisons We perform evaluations on three different datasets: Our high-quality studio dataset, a publicly available studio dataset (Facescape [69]), and in-the-wild captures from a mobile and a cellphone camera. For the studio datasets, we assume calibrated cameras. For the in-the-wild captures, we estimate camera parameters with Mediapipe [28]. The metrics for the quantitative comparisons are computed after cropping the images to squares and setting the background to black with foreground masks from a state-of-the-art network [38]. For more details, please refer to the supplementary material.

6. Results

We perform extensive evaluation and experiments to demonstrate i) our core claims - high resolution, few shot, in-the-wild synthesis, ii) improved performance over the state-of-the-art methods, iii) ablation of various design choices. We also encourage the reader to see the video results and more insightful evaluations in the supp. mat.

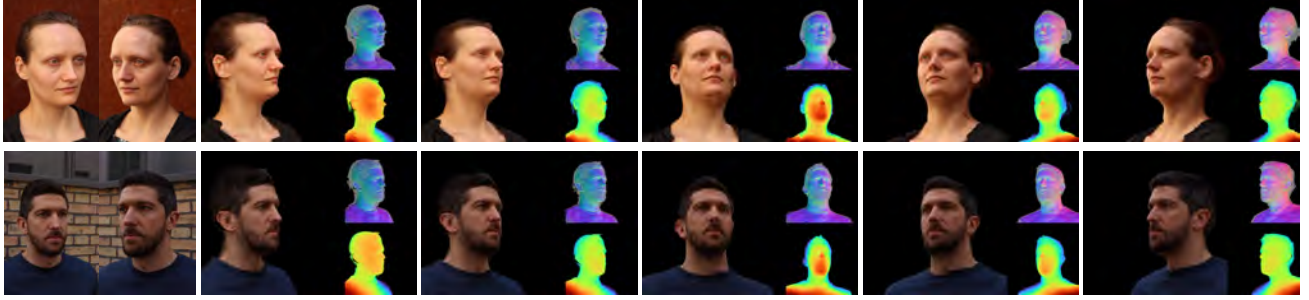


Figure 7. In-the-wild Results. We reconstruct a target identity from two images acquired with a consumer camera (left). Note how the novel views can extrapolate from the input camera angles. The inlays show the normals (top) and depth (bottom). The hair density is low, thus the grey normal colour in that region. We encourage the reader to see the supp. mat. for the high-resolution results and videos.

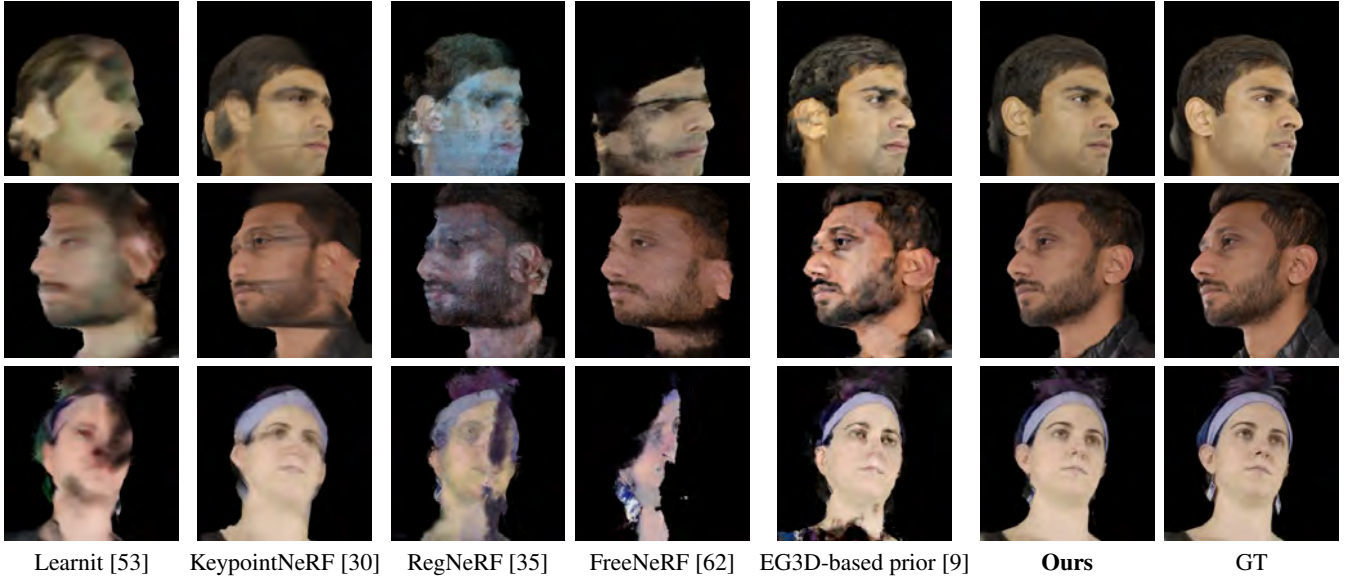


Figure 8. Visual comparison when given two target views. Our method consistently produces more pleasing results. Please see Tbl. 1 for metrics and the supplementary material for implementation details and results on more than two target views.

| Initialization | PSNR \uparrow | SSIM \uparrow | LPIPS \downarrow |
|---------------------------|-----------------|-----------------|--------------------|
| Furthest | 23.91 | 0.7876 | 0.2041 |
| Nearest | 24.41 | 0.7900 | 0.2002 |
| Mean | 24.61 | 0.7934 | 0.1959 |
| Noise | 24.66 | 0.7957 | 0.1998 |
| Zeros | 24.65 | 0.7941 | 0.1944 |
| Inversion (Ours) | 25.69 | 0.8040 | 0.1905 |

Table 2. Ablation on various types of initialising $\mathbf{w}_{\text{target}}$ when finetuning the model. We compare taking the mean across all latent codes during training; initialising it with zeros, Gaussian noise; and copying the latent code of the nearest or furthest neighbor in the training set. Inversion (**Ours**) performs best. Please refer to the supplementary material for visual examples.

6.1. Ultra-high Resolution Synthesis

We demonstrate ultra high resolution synthesis after finetuning our 512×768 prior model to sparse high-resolution images in the studio setting (Fig. 1) and in-the-wild (Fig. 7).

4K Novel Views from Three Views Figure 1 shows 4096×4096 (4K) renderings after finetuning to three views of a held-out subject from our studio dataset. Note the range of the rendered novel views and the quality of synthesis results for such an out-of-distribution test subject at 4K resolution. From just three images, our method learns a highly detailed and photorealistic volumetric model of a face. We synthesise smooth and 3D consistent camera trajectories while preserving challenging details such as individual hair strands, skin pores and eyelashes. Our model learns both consistent geometry and fine details of individual hair strands and microgeometry of the skin, making the synthesised images barely distinguishable from captured views. Please see the supplementary material for video results and results on other subjects.

2K Novel Views from Two in-the-wild Views Our method also affords reconstruction from in-the-wild cap-

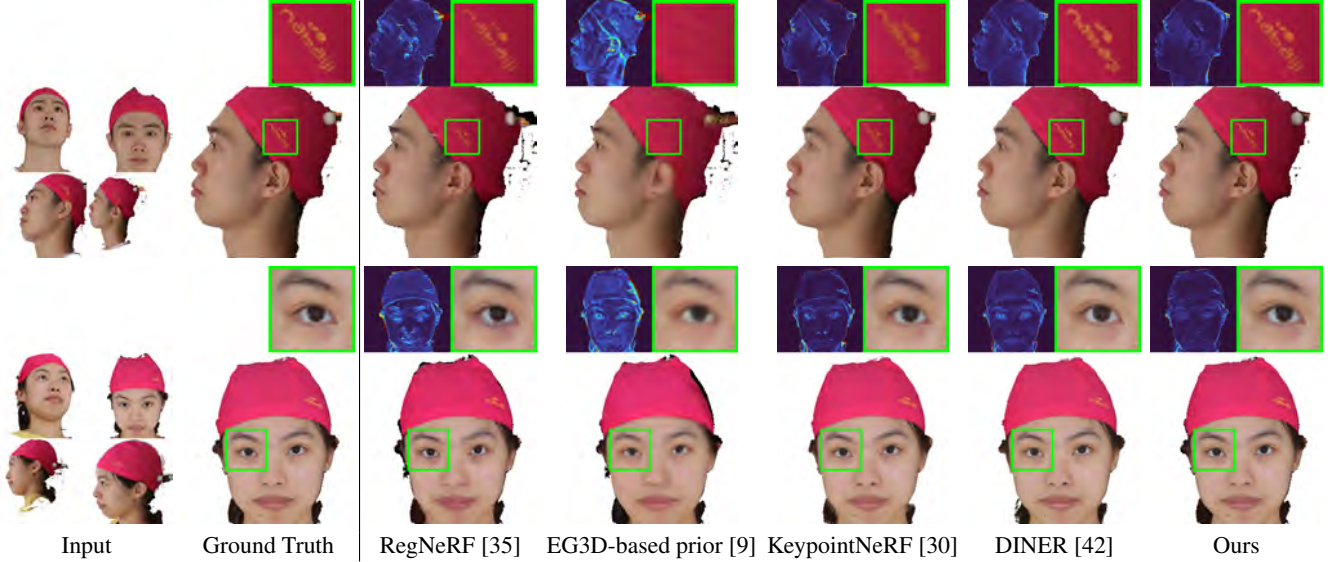


Figure 9. Comparison with the state-of-the-art on holdout identities from FaceScape [69]. Each method is given four input views and we show novel views and the L1 residue. Please see the supp. mat. for implementation details, more examples, and detailed metrics.

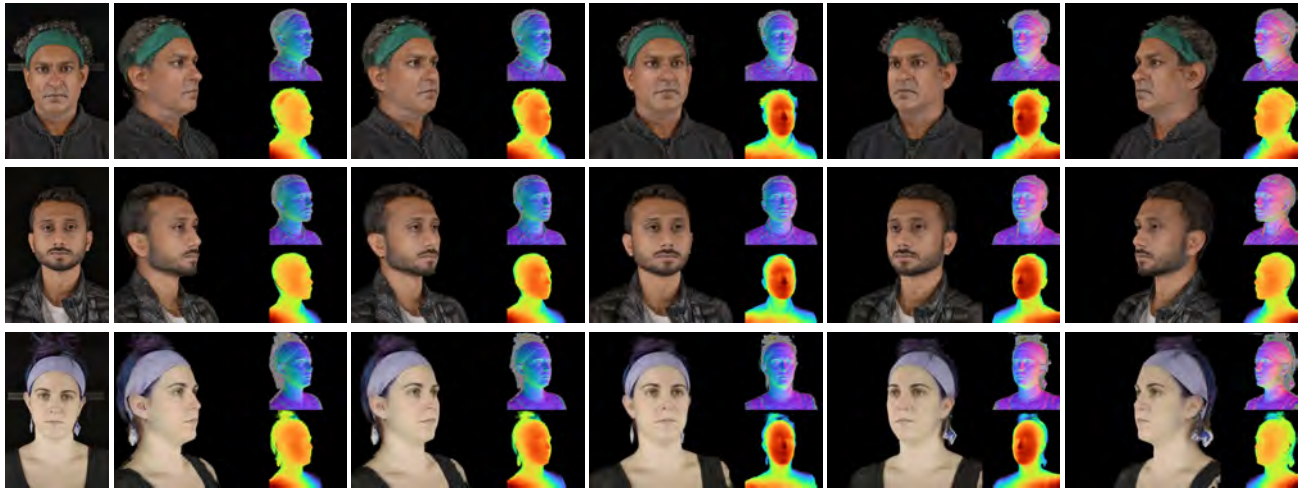


Figure 10. Single Image Reconstruction Results. From left to right: input image captured using a studio setup, synthesised views around the subject face using a single frontal view for model fitting.

tures from a single camera. We use a digital camera to capture two images. Results are shown in Fig. 7. The upper row was captured outdoors in front of a wall; the bottom row was captured in a room. Please see the supplementary material for more examples and videos.

6.2. Comparison with Related Work

Our goal is high-resolution novel view synthesis from sparse inputs. We perform comparisons by training related works [9, 30, 35, 53, 62] on our studio dataset and rendering results for unseen views at resolution 1024×1024 (1K). Since the task of novel view synthesis becomes substantially easier when given more views of the target subject, we

perform comparisons for different number of views ranging from two to seven. Fig. 8 and Tbl. 1 show that our method can handle difficult cases at high resolution and clearly outperforms all related works when reconstructing from two views. Please see the supp. mat. for results on more views.

We observe that some of the related methods perform significantly better at lower resolutions and when given more than just two views of the target subject. Hence, we complement our comparisons with a comparison on the FaceScape dataset [69]. We follow the setting of the best performing related work, DINER [42], and use four reference images at resolution 256×256 . Fig. 9 displays visuals and the supplementary document provides metrics. Note

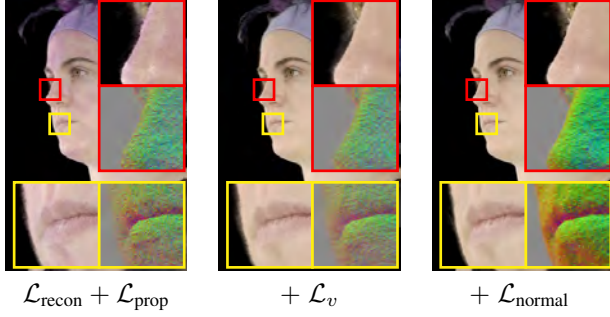


Figure 11. Ablation on the choice of regularisers. Without any regularisation, the view branch of the model overfits to the view direction from the sparse input signal. Additional regularisers allows the model to fit to a target identity from very sparse views.

that KeypointNerf [30] and DINER [42] were trained on Facescape while ours is not. This means that our scores represent results in the "out-of-distribution" setting.

6.3. Single Image Fitting

Our method is also capable of fitting to a single image and still produces detailed results. We show such result on held-out test subjects from our dataset in Fig. 10. Note the consistent depth and normal maps and photorealistic renderings. This indicates that our model learns a strong prior over head geometry which helps it resolve depth ambiguity to reconstruct a cohesive density field for the head, including challenging regions like hair.

6.4. Ablations

Initialisation The initialisation of the latent code plays a key role in achieving good results. We ablate various initialisation choices such as: i) a zero vector, ii) Gaussian noise, iii) the mean over the training latent codes, iv) the nearest and furthest neighbour in the training set defined by a precomputed embedding [49], and v) inversion (Ours). We finetune the prior model to two views of three holdout identities and report the results in Tbl. 2. Inversion performs best in all metrics.

Regularisation We also ablate the choice of regularisation for the model finetuning. Fig. 11 shows that without any regularisation, the view branch of the model overfits to the view direction from the sparse input signal. We observe that the parameter weights of the view branch become very large and dominate the colour observed from a particular view. To mitigate this, we regularise the L2 norm of the weights using \mathcal{L}_v (green highlight in Fig. 5). However, the model still overfits by generating a fuzzy surface that produces highly specular effects from the optimised views but has incorrect geometry. To regularise the geometry, we extend the trunk of our model with a branch predicting normal



Figure 12. We show results for challenging lighting conditions with shadows and specular reflections, e.g., on the forehead. The right column lists PSNR, SSIM, and LPIPS.

and supervise it with the analytical normals [55]. With both regularisation terms, the model can be robustly fit to a target identity from very sparse views.

Challenging Lighting Conditions Our method can generate high-quality novel views even under challenging lighting conditions with shadows and specular reflections, see Fig. 12.

Further Ablations We perform further ablations for fitting to a higher number of target views, for different configurations of our prior models, and for frozen latent codes during model finetuning. Please see the supplementary material for results.

7. Conclusion

We present a method that can create ultra high-resolution NeRFs of unseen subjects from as few as two images, yielding quality that surpasses other state-of-the-art methods. While our method generalises well along several dimensions such as identity, resolution, viewpoint, and lighting, it is also impacted by the limitations of our dataset. While minor deviations from a neutral expression such as smiles can be synthesised, it struggles with extreme expressions. Clothing and accessories are also harder to synthesise. We show examples of such failure cases in the supplementary. Our model fitting process can take a considerable amount of time, particularly at higher resolutions. While some of these problems can be solved with more diverse data, others are excellent avenues for future work.

Acknowledgments. We thank Emre Aksan for insightful discussions and Malte Prinzler for sharing DINER results.

References

- [1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.
- [3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [4] Pascal Bérard, Derek Bradley, Maurizio Nitti, Thabo Beeler, and Markus Gross. High-quality capture of eyes. *ACM Trans. Graph.*, 33(6), nov 2014.
- [5] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2640–3498, 2018.
- [6] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022.
- [7] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020.
- [8] 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. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [9] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021.
- [10] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM transactions on graphics*, 2021.
- [11] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.
- [12] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [13] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
- [14] Jose I. Echevarria, Derek Bradley, Diego Gutierrez, and Thabo Beeler. Capturing and stylizing hair for 3d fabrication. *ACM Trans. Graph.*, 33(4), jul 2014.
- [15] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.
- [16] P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Perez, T. Beeler, and C. Theobalt. Corrective 3D Reconstruction of Lips from Monocular Video. *ACM Transactions on Graphics (TOG)*, 35(6), 2016.
- [17] Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Practical dynamic facial appearance modeling and acquisition. *ACM Trans. Graph.*, 37(6), dec 2018.
- [18] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.
- [19] Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *Technical Report*, 2023.
- [20] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021.
- [21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.
- [22] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.
- [23] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022.
- [24] Gengyan Li, Abhimitra Meka, Franziska Mueller, Marcel C Buehler, Otmar Hilliges, and Thabo Beeler. Eyenerf: a hybrid representation for photorealistic synthesis, animation and relighting of human eyes. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.
- [25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- [26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019.
- [27] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), jul 2021.
- [28] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chu-

- Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.
- [29] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019.
- [30] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, 2022.
- [31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.
- [32] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.
- [33] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.
- [34] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.
- [35] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [36] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.
- [37] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *arXiv e-prints*, pages arXiv–2112, 2021.
- [38] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021.
- [39] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019.
- [40] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.
- [41] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021.
- [42] Malte Prinzler, Otmar Hilliges, and Justus Thies. Diner: Depth-aware image-based neural radiance fields, 2022.
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021.
- [44] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021.
- [45] 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. Vorf: Volumetric relightable faces. *British Machine Vision Conference (BMVC)*, 2022.
- [46] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolerf: Learn from one look. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1558–1567, 2022.
- [47] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021.
- [48] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022.
- [49] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [50] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.
- [51] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020.
- [52] Feitong Tan, Sean Fanello, Abhimitha Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH ’22, New York, NY, USA, 2022. Association for Computing Machinery.
- [53] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021.
- [54] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from

- sparse and noisy poses. *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2023.
- [55] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.
 - [56] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *arXiv preprint arXiv:2111.13260*, 2021.
 - [57] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.
 - [58] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021.
 - [59] C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. Gross, and T. Beeler. Model-Based Teeth Reconstruction. *ACM Transactions on Graphics (TOG)*, 35(6), 2016.
 - [60] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. *arXiv preprint arXiv:2206.07255*, 2022.
 - [61] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 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*, pages 736–753. Springer, 2022.
 - [62] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.
 - [63] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021.
 - [64] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
 - [65] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. *arXiv preprint arXiv:2208.05751*, 2022.
 - [66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *CoRR*, abs/2010.07492, 2020.
 - [67] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
 - [68] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. CIPS-3D: A 3D-Aware Generator of GANs Based on Conditionally-Independent Pixel Synthesis. 2021.
 - [69] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *arXiv preprint arXiv:2111.01082*, 2021.