DINER: Depth-aware Image-based NEural Radiance fields

Malte Prinzler[1,3]
malte.prinzler@tuebingen.mpg.de

Otmar Hilliges[2]
otmar.hilliges@inf.ethz.ch

Justus Thies[1]
justus.thies@tuebingen.mpg.de

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]ETH Zürich [3]Max Planck ETH Center for Learning Systems

Figure 1. Based on sparse input views, we predict depth and feature maps to infer a volumetric scene representation in terms of a radiance field which enables novel viewpoint synthesis. The depth information allows us to use input views with high relative distance such that the scene can be captured more completely and with higher synthesis quality compared to previous state-of-the-art methods.

Abstract

We present Depth-aware Image-based NEural Radiance fields (DINER)[1]*. Given a sparse set of RGB input views, we predict depth and feature maps to guide the reconstruction of a volumetric scene representation that allows us to render 3D objects under novel views. Specifically, we propose novel techniques to incorporate depth information into feature fusion and efficient scene sampling. In comparison to the previous state of the art, DINER achieves higher synthesis quality and can process input views with greater disparity. This allows us to capture scenes more completely without changing capturing hardware requirements and ultimately enables larger viewpoint changes during novel view synthesis. We evaluate our method by synthesizing novel views, both for human heads and for general objects, and observe significantly improved qualitative results and increased perceptual metrics compared to the previous state of the art.*

[1]https://malteprinzler.github.io/projects/diner/diner.html

1. Introduction

In the past few years, we have seen immense progress in digitizing humans for virtual and augmented reality applications. Especially with the introduction of neural rendering and neural scene representations [43, 44], we see 3D digital humans that can be rendered under novel views while being controlled via face and body tracking [3, 11, 14, 15, 24, 32, 37, 47, 50, 51, 57]. Another line of research reproduces general 3D objects from few input images without aiming for control over expressions and poses [6, 25, 38, 41, 46, 55]. We argue that this offers significant advantages in real-world applications like video-conferencing with holographic displays: (i) it is not limited to heads and bodies but can also reproduce objects that humans interact with, (ii) even for unseen extreme expressions, fine texture details can be synthesized since they can be transferred from the input images, (iii) only little capturing hardware is required e.g. four webcams suffice, and (iv) the approach can generalize across identities such that new participants could join the conference ad hoc without requiring subject-specific optimization. Because of these advantages, we study the scenario of

reconstructing a volumetric scene representation for novel view synthesis from sparse camera views. Specifically, we assume an input of four cameras with high relative distances to observe large parts of the scene. Based on these images, we condition a neural radiance field [27] which can be rendered under novel views including view-dependent effects. We refer to this approach as *image-based neural radiance fields*. It implicitly requires estimating the scene geometry from the source images. However, we observe that even for current state-of-the-art methods the geometry estimation often fails and significant synthesis artifacts occur when the distance between the source cameras becomes large – which conflicts with the goal of modeling the target scene more completely while keeping the capture setup as simple as possible. Specifically, current models fail in this challenging scenario because they rely on implicit correspondences between the different views. Recent research demonstrates the benefits of exploiting triangulated landmarks to guide the correspondence search [25]. However, landmarks have several drawbacks: They only provide sparse guidance, are limited to specific classes, hence, cannot represent arbitrary objects beyond faces or bodies, and the downstream task is bounded by the quality of the keypoint estimation, which is known to depend on the viewpoint, e.g., landmark detectors produce noisy results for profile views.

To this end, we propose *DINER* to compute an image-based neural radiance field that is guided by estimated dense depth. This has significant advantages: depth maps are not restricted to specific object categories (like human faces or bodies), provide dense guidance, and are easy to attain via either a commodity depth sensor or off-the-shelf depth estimation methods. Specifically, we leverage a state-of-the-art depth estimation network [8] to predict depth maps for each of the source views and employ an encoder network that regresses pixel-aligned feature maps. DINER exploits the depth maps in two important ways: (i) we condition the neural radiance field on the deviation between sample location and depth estimates which provides strong prior information about visual opacity, and (ii) depth maps allow us to place samples close to the estimated surfaces which increases the sample density in regions that actually contribute to the synthesis result. Furthermore, we improve the extrapolation capabilities of image-based NeRFs by padding and positionally encoding the input images before applying the feature extractor. Our model is trained on many different scenes and at inference time, four input images suffice to reconstruct the target scene in one inference step without requiring scene-specific optimization. As a result, compared to the previous state of the art, *DINER* can reconstruct 3D scenes from more distinct source views with better visual quality, while allowing for larger viewpoint changes during novel view synthesis. We evaluate our method on the large-scale FaceScape dataset [53] on the task of novel view synthesis for human heads from only four highly diverse source views. Since our method is not bound to a specific object category, we also evaluate our model on the task of novel view synthesis for general objects on the DTU dataset [18]. For both datasets, our model outperforms all baselines by a significant margin.

In summary, *DINER* is a novel method that produces volumetric scene reconstructions from few source views with higher quality and completeness than the previous state of the art. In summary, we contribute:

- an effective approach to condition image-based NeRFs on depth maps predicted from the RGB input,
- a novel depth-guided sampling strategy that increases efficiency,
- and a method to improve the extrapolation capabilities of image-based NeRFs by padding and positionally encoding the source images prior to feature extraction.

2. Related Work

Our work is related to recent approaches on 3D head avatar generation, and to general neural radiance fields that are reconstructed from a sparse set of input images. In the following, we will give a brief overview of related works and detail the fundamentals of our approach in Section 3.

Neural Radiance Fields Neural radiance fields (NeRF) [27] and their derivatives have become a popular choice for representing photo-realistic 3D scenes, both for static [1, 4, 22, 26, 28, 35, 42, 52] and dynamic [11, 19, 30, 31, 33] cases. While originally requiring many training images per scene and long training and inference times, latest research focused on increasing data sufficiency [7, 29, 35, 36] as well as making NeRFs more efficient and faster [4, 5, 13, 17, 28]. Some of these methods even exploit depth information [7, 35, 36], however, they only assume that depth maps are given for target views at training time. Our method instead exploits depth maps of the source views which do not necessarily coincide with the target views and which are predicted by a depth estimator both at training and test time.

Image-Based Rendering The above methods require scene-specific optimization which is time-consuming and expensive. Methods that extract features from as little as one source image and warp them for novel view synthesis [9, 39, 48, 56] offer a fast and efficient alternative but produce artifacts for strong rotations. Image-based NeRFs [6, 25, 46, 55] solve such artifacts through explicit 3D reasoning by using few source images to reconstruct NeRF-like representations in one inference step. However, for source images with small overlap artifacts occur. Triangulated landmarks may be used to guide the geometry esti-

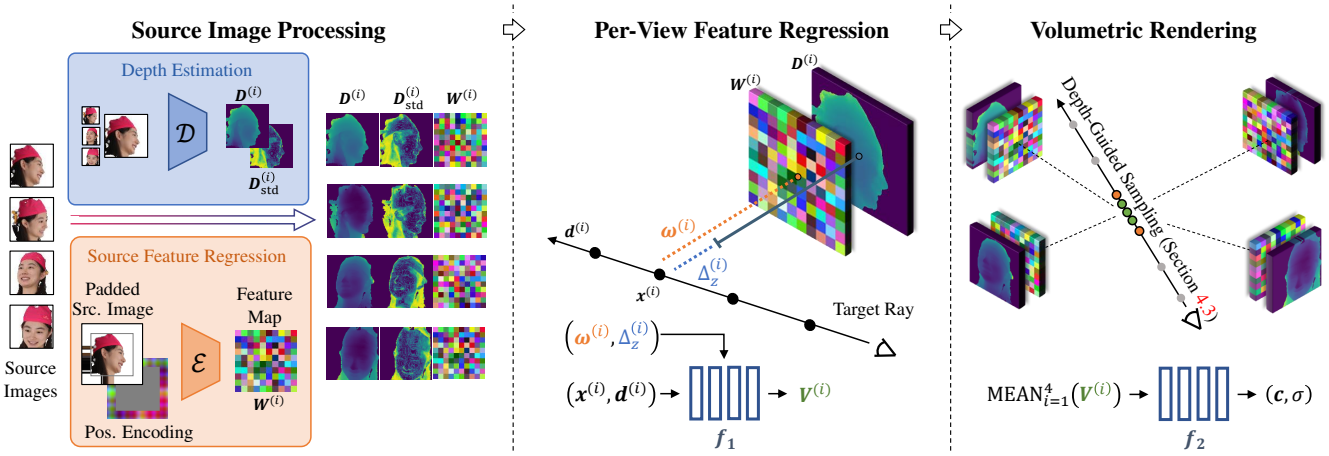

Figure 2. Method overview. Given few source images, we first regress depth and feature maps. Image padding and positional encoding prior to the feature map prediction improve extrapolation beyond the source image frustums (Section 4.2). Sampled points along target rays are projected onto the source camera planes to interpolate feature vectors $\boldsymbol{\omega}^{(i)}$ and the deviations $\Delta_z^{(i)}$ between the sampling point and predicted depth. They are processed into view-wise intermediate feature vectors $\boldsymbol{V}^{(i)}$ (Section 4.1). The average-pooled intermediate feature vectors of all source views determine the color $\boldsymbol{c}$ and opacity σ of each sampling point. The final colors of the target rays are obtained through standard volumetric rendering (1). Depth-guided sampling increases sampling efficiency (Section 4.3).

mation [25], but they only offer sparse guidance and cannot be applied to arbitrary objects. Our method instead exploits depth maps which are more general and provide denser guidance.

Head Avatars The research on head avatar synthesis can be separated along many dimensions [43, 58]. With regard to the choice of representation, the different approaches either rely on image-to-image translation [2, 21, 39, 45, 56], volumetric rendering [11, 12, 16, 24, 37, 49, 50], or non-transparent surface textures [3, 10, 14, 20, 34, 57]. The methods reconstruct head avatars from as little as one source image [2, 9, 16, 20, 39, 48, 56], monocular videos [3, 11, 14, 21], or entire camera domes [12, 23, 24, 49, 50]. Our method relies on four source images to reconstruct a volumetric head representation, allowing us to synthesize even thin semi-transparent structures like hair plausibly. The focus of our method is not to learn an animatable avatar, instead, we focus on reconstructing a high-quality volumetric representation from few (e.g., four) and widely spaced source views to enable light-weight capture of immersive, viewpoint-independent video (e.g., for 3D telepresence applications). Furthermore, our approach is not limited to human heads but can be applied to general objects as well.

3. Background

Before detailing the architecture of our pipeline, we briefly review the concepts which it is based on, namely NeRFs in general and image-based NeRFs in particular, and introduce the used notation.

NeRF Neural radiance fields (NeRF) [27] employ multi-layer perceptrons (MLP) to implicitly parameterize a continuous volumetric function f that maps a 3D position $\boldsymbol{x}$ and a view direction vector $\boldsymbol{d}$ to a view-dependent color value $\boldsymbol{c}$ and an isotropic optical density σ, so that $(\boldsymbol{c}, \sigma) = f(\boldsymbol{x}, \boldsymbol{d})$. To render the scene, rays are cast into the scene for every pixel of the target image. We assume such a ray is parametrized by $\boldsymbol{r}(t) = \boldsymbol{o} + t \cdot \boldsymbol{d}$ with near and far plane $t_{\text{near}}, t_{\text{far}}$ respectively. NeRF samples 3D points along the ray $t_j \sim [t_{\text{near}}, t_{\text{far}}]$, estimates their color and optical density $(\boldsymbol{c}_j, \sigma_j) = f(\boldsymbol{r}(t_j), \boldsymbol{d})$, and integrates the results along the ray following volumetric rendering:

$$\hat{\boldsymbol{C}}(\boldsymbol{r}) = \sum_{i=j}^{N} T_j (1 - \exp(-\sigma_j \delta_j)) \boldsymbol{c_j} \text{ with } T_j = \exp\left(- \sum_{k=1}^{j-1} \sigma_k \delta_k \right), \quad (1)$$

where $\delta_j = t_{j+1} - t_j$ denotes the distance between adjacent samples. In practice, NeRF employs a coarse-to-fine strategy to place samples more efficiently: a coarse MLP is queried to determine regions of importance around which the sample density is increased for evaluating a fine MLP that regresses the final ray color.

Image-Based NeRFs Image-based NeRFs enable generalization across scenes by conditioning the NeRF on features extracted from source images. We build our model on top of the pixelNeRF [55] pipeline. Assuming N source images $\{\boldsymbol{I}^{(i)}\}_{i=1}^{N}$ with known extrinsics $\boldsymbol{P}^{(i)} = [\boldsymbol{R}^{(i)} \; \boldsymbol{t}^{(i)}]$ and intrinsics $\boldsymbol{K}^{(i)}$, a 2D-convolutional encoding network $\mathcal{E}$ extracts feature maps $\boldsymbol{W}^{(i)}$ for each source image:

$$\boldsymbol{W}^{(i)} = \mathcal{E}(\boldsymbol{I}^{(i)}). \quad (2)$$

For obtaining the color and opacity of a point, its 3D position $\boldsymbol{x}$ and view direction $\boldsymbol{d}$ are first transformed into the source view coordinate systems:

$$\boldsymbol{x}^{(i)} = \boldsymbol{P}^{(i)} \circ \boldsymbol{x}, \quad \boldsymbol{d}^{(i)} = \boldsymbol{R}^{(i)} \circ \boldsymbol{d}, \tag{3}$$

after which $\boldsymbol{x}^{(i)}$ is projected onto the respective feature map to sample a feature vector $\boldsymbol{\omega}^{(i)}$ trough bilinear interpolation:

$$\boldsymbol{\omega}^{(i)} = \boldsymbol{W}^{(i)} \left(\boldsymbol{K}^{(i)} \circ \boldsymbol{x}^{(i)} \right). \tag{4}$$

The NeRF MLP f is split into two parts, f_1 and f_2. f_1 processes the input coordinates of the sampling point alongside the sampled feature vectors into intermediate feature vectors $\boldsymbol{V}^{(i)}$ for every view independently:

$$\boldsymbol{V}^{(i)} = f_1 \left(\boldsymbol{x}^{(i)}, \boldsymbol{d}^{(i)}, \boldsymbol{\omega}^{(i)} \right). \tag{5}$$

The feature vectors from different views are aggregated through average pooling and then processed by f_2 to regress the final color and density values:

$$\boldsymbol{c}, \sigma = f_2 \left(\text{mean}_i \left\{ \boldsymbol{V}^{(i)} \right\} \right). \tag{6}$$

During training, the l_1 distance between estimated ray colors and ground truth RGB values is minimized.

4. Method

Given a sparse set of input images $\{\boldsymbol{I}^{(i)}\}$ $(i = 1...N = 4)$, our approach infers a NeRF which allows rendering novel views of the scene. We estimate the depth in the given input views and propose two novel techniques to leverage this information during scene reconstruction: (i) the NeRF MLP is conditioned on the difference between sample location and estimated depth which serves as a strong prior for the visual opacity (Section 4.1), and (ii) we focus the scene sampling on the estimated surface regions, i.e. on regions that actually contribute to the scene appearance (Section 4.3). Furthermore, we propose to pad and positionally encode the source images prior to the feature extraction to improve extrapolation capabilities beyond the source view frustums (Section 4.2). Please refer to Figure 2 for an overview of our pipeline.

4.1. Depth Conditioning

Our method is based on pixelNeRF (see Section 3) and leverages the attention-based TransMVSNet [8] architecture to estimate depth from the sparse observations. TransMVSNet takes all four input images $\boldsymbol{I}^{(i)}$ as input and predicts the per-view depth maps $\boldsymbol{D}^{(i)}$ as well as the corresponding standard deviations $\boldsymbol{D}_{\text{std}}{}^{(i)}$. For each sampling point $\boldsymbol{x}$, we calculate the difference $\Delta_z^{(i)}$ between its z-component in the i-th camera space and its corresponding projected depth value: $\Delta_z^{(i)} = \boldsymbol{D}^{(i)} \left(\boldsymbol{K}^{(i)} \circ \boldsymbol{x}^{(i)} \right) - \boldsymbol{x}_{[z]}^{(i)}$, which is input to the pixelNeRF MLP f_1 as additional conditioning:

$$\boldsymbol{V}^{(i)} = f_1 \left(\boldsymbol{x}^{(i)}, \boldsymbol{d}^{(i)}, \boldsymbol{\omega}^{(i)}, \gamma \left(\Delta_z^{(i)} \right) \right). \tag{7}$$

$\gamma(\cdot)$ denotes positional encoding with exponential frequencies as proposed in [27].

4.2. Source Feature Extrapolation

When projecting sampling points on the source feature maps, image-based NeRFs typically apply border padding, i.e. points outside the map's boundaries are assigned constant feature vectors irrespective of their distance to the feature map. During synthesis, this causes smearing artifacts in regions that are not visible by the source images (see Figure 5, third column). To solve this, we make two modifications to the source images $\boldsymbol{I}^{(i)}$ *before* applying the encoder network $\mathcal{E}$. We apply border padding and add channels with positional encodings of the padded pixel coordinates, resulting in $\boldsymbol{I'}^{(i)} = concatenate \left(\boldsymbol{I}_{\text{pad}}^{(i)}, \boldsymbol{\Gamma} \right)$, where $\boldsymbol{\Gamma}$ contains the pixel-wise positional encodings for the padded regions:

$$\boldsymbol{\Gamma}_{[u,v]} = \begin{cases} \gamma(u,v) & \text{if } (u,v) \notin \boldsymbol{I}^{(i)} \\ 0 & \text{if } (u,v) \in \boldsymbol{I}^{(i)}. \end{cases} \tag{8}$$

The positional encoding supports $\mathcal{E}$ in regressing distinctive features in padded regions where the extrapolated color values are constant.

4.3. Depth-Guided Sampling

Since only object surfaces and the immediate surroundings contribute to ray colors, we aim to focus our sampling on these regions. The estimated depth maps provide strong priors about the placements of such surfaces. This allows us to increase the sample density in relevant regions which improves the synthesis quality (see Section 5.2). Figure 3 provides an overview of our approach. Note that while previous work incorporates depth information of target views during NeRF training [7, 35, 36], we are the first to exploit depth from input views that do not coincide with the target view and which we predict both at training and test time.

Depth-Guided Probability Fields For each input image, the depth estimator provides pixel-wise depth expectation values $\boldsymbol{D}^{(i)}$ and standard deviations $\boldsymbol{D}_{\text{std}}{}^{(i)}$. These maps define pixel-aligned probability density fields for the presence of object surfaces. Assuming a ray $\boldsymbol{r}(t)$ with near plane t_{near} and far plane t_{far}, we first uniformly sample a large set of N_{cand} candidate samples along the ray :

$$\{\boldsymbol{x}\} = \{\boldsymbol{r}(t) \mid t \sim [t_{\text{near}}, t_{\text{far}}]\}. \tag{9}$$

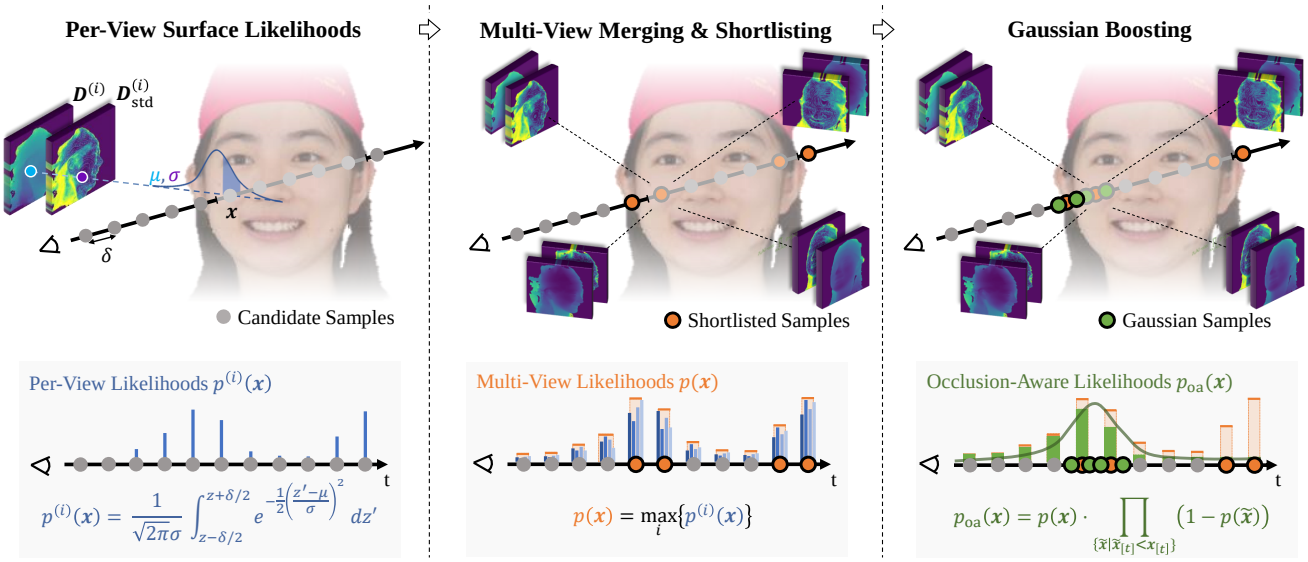

Figure 3. Depth-guided sampling. We sample candidate points along the target ray and evaluate their surface likelihoods given the depth estimates for each input view. The view-wise likelihoods are aggregated through max-pooling and we shortlist the most-likely samples. Additional points are sampled according to a Gaussian distribution that was fitted against the occlusion-aware likelihoods of all candidates.

For each of the input views, we project $\{\boldsymbol{x}\}$ onto the respective depth maps and determine the likelihood of $\boldsymbol{x}$ being a surface point:

$$p^{(i)}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}\sigma} \int_{z-\delta/2}^{z+\delta/2} e^{-\frac{1}{2}\left(\frac{z'-\mu}{\sigma}\right)^2} dz', \tag{10}$$

$$\text{with} \quad \mu = \boldsymbol{D}^{(i)}\left(\boldsymbol{K}^{(i)} \circ \boldsymbol{x}^{(i)}\right), \quad z = \boldsymbol{x}^{(i)}_{[z]},$$
$$\sigma = \boldsymbol{D}^{(i)}_{\text{std}}\left(\boldsymbol{K}^{(i)} \circ \boldsymbol{x}^{(i)}\right), \quad \delta = (t_{\text{far}} - t_{\text{near}})/N_{\text{cand}}.$$

We perform backface culling by first calculating normals from each depth map and then discarding samples if the angle between the ray direction and the projected normal value is smaller than $90°$. The likelihood $p(\boldsymbol{x})$ of a point coinciding with a surface is determined by view-wise max-pooling:

$$p(\boldsymbol{x}) = \max_i \left\{ p^{(i)}(\boldsymbol{x}) \right\}. \tag{11}$$

We shortlist N_{samples} candidates with the highest likelihoods for sampling the NeRF.

Gaussian Boosting To further improve sampling efficiency, we sample additional points around the termination expectation value of the ray. The occlusion-aware likelihoods of the ray $\boldsymbol{r}$ terminating in sample $\boldsymbol{x}$ is given by:

$$p_{\text{oa}}(\boldsymbol{x}) = p(\boldsymbol{x}) \cdot \prod_{\{\tilde{\boldsymbol{x}} \mid \tilde{\boldsymbol{x}}_{[t]} < \boldsymbol{x}_{[t]}\}} (1 - p(\tilde{\boldsymbol{x}})). \tag{12}$$

Please note the simplified notation $\boldsymbol{x}_{[t]} := t$ so that $\boldsymbol{r}(t) = \boldsymbol{x}$. We fit a Gaussian distribution against the occlusion-aware likelihoods along the ray (see Figure 3 right) and sample N_{gauss} points from it which are added to the shortlisted candidates.

4.4. Loss Formulation

Our loss formulation consists of a per-pixel reconstruction error $\mathcal{L}_{l_1}$ as well as a perceptual loss $\mathcal{L}_{\text{vgg}}$ [40]. While a perceptual loss improves high-frequency details, it can introduce color shifts (see Figure 5). Thus, we introduce an anti-bias term $\mathcal{L}_{\text{ab}}$ which corresponds to a standard l_1 loss that is applied to downsampled versions of prediction and ground truth. $\mathcal{L}_{\text{ab}}$ effectively eliminates color shifts while being robust against minor misalignments, hence in contrast to the standard l_1 loss, it does not introduce low-frequency bias. Let $\boldsymbol{P}$ denote the ground truth image patch and $\hat{\boldsymbol{P}}$ its predicted counterpart, then $\mathcal{L}_{\text{ab}}$ is given by:

$$\mathcal{L}_{\text{ab}} = \left|\left| \text{DS}_k(\boldsymbol{P}) - \text{DS}_k(\hat{\boldsymbol{P}}) \right|\right|_1^1, \tag{13}$$

where $\text{DS}_k(\cdot)$ denotes k-fold downsampling. Our full objective function is defined as:

$$\mathcal{L} = w_{l_1} \cdot \mathcal{L}_{l_1} + w_{\text{vgg}} \cdot \mathcal{L}_{\text{vgg}} + w_{\text{ab}} \cdot \mathcal{L}_{\text{ab}}, \tag{14}$$

where $\mathcal{L}_{l_1}$ corresponds to a pixel-wise l_1 distance. We use an Adam optimizer with standard parameters and a learning rate of 10^{-4} and train on patches of 64×64 px on a single NVIDIA A100-SXM-80GB GPU with batch size 4 for 330k iterations which takes 4 days. Note that depth-guided sampling allows reducing the samples per ray by a factor of 4 w.r.t. pixelNeRF because samples are focused on non-empty areas which improves the performance (see Table 2).

5. Results

We conduct the validation of our approach on the FaceScape dataset [53] which contains more than 400k por-

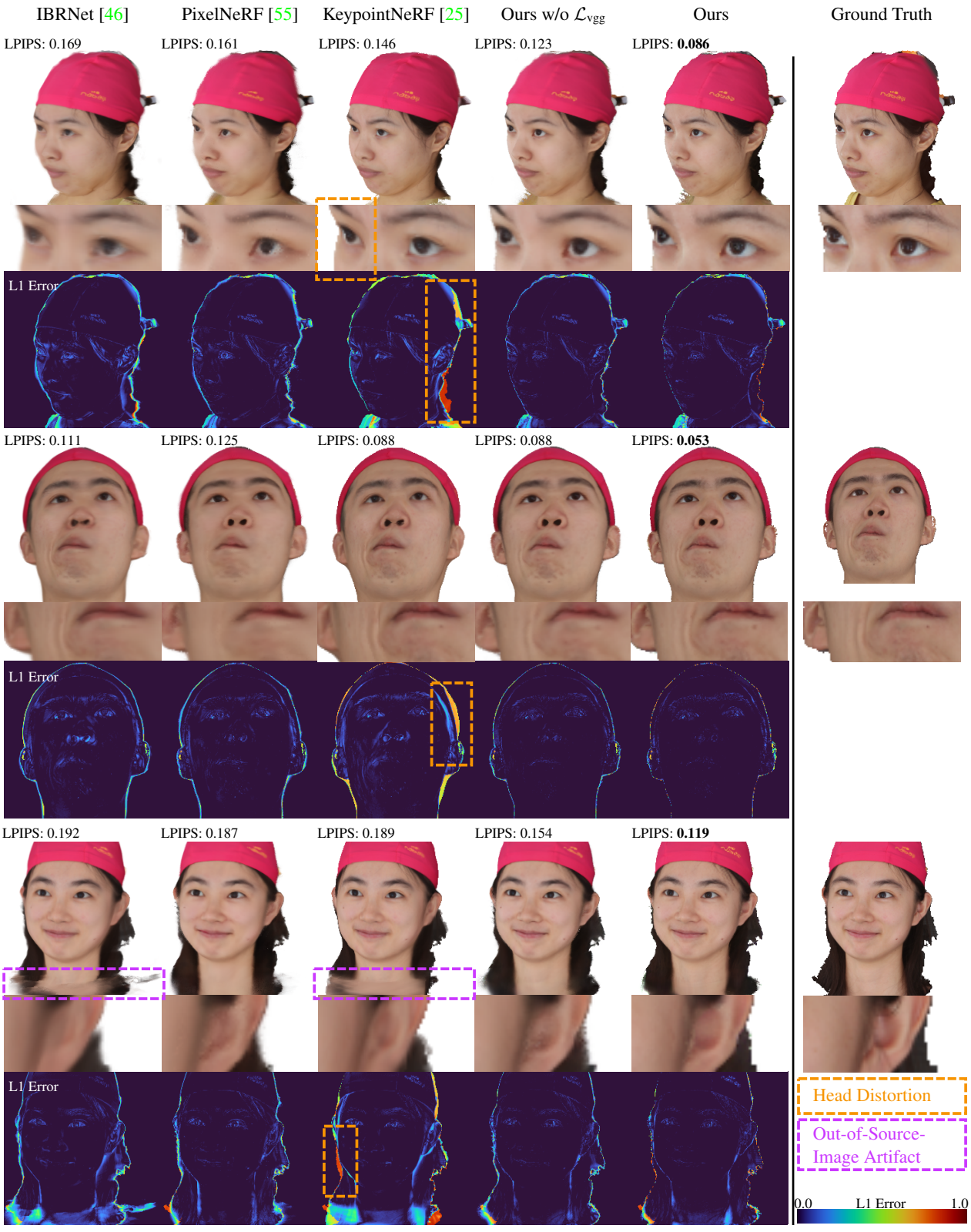

Figure 4. Qualitative comparison on FaceScape. IBRNet and KeypointNeRF produce artifacts for regions outside of the source images (pink). pixelNeRF handles these aspects better but produces blurry results. KeypointNeRF synthesizes heads with deformations (orange). Even without the perceptual loss, our method yields better results and adding the perceptual loss further emphasizes high-frequency details.

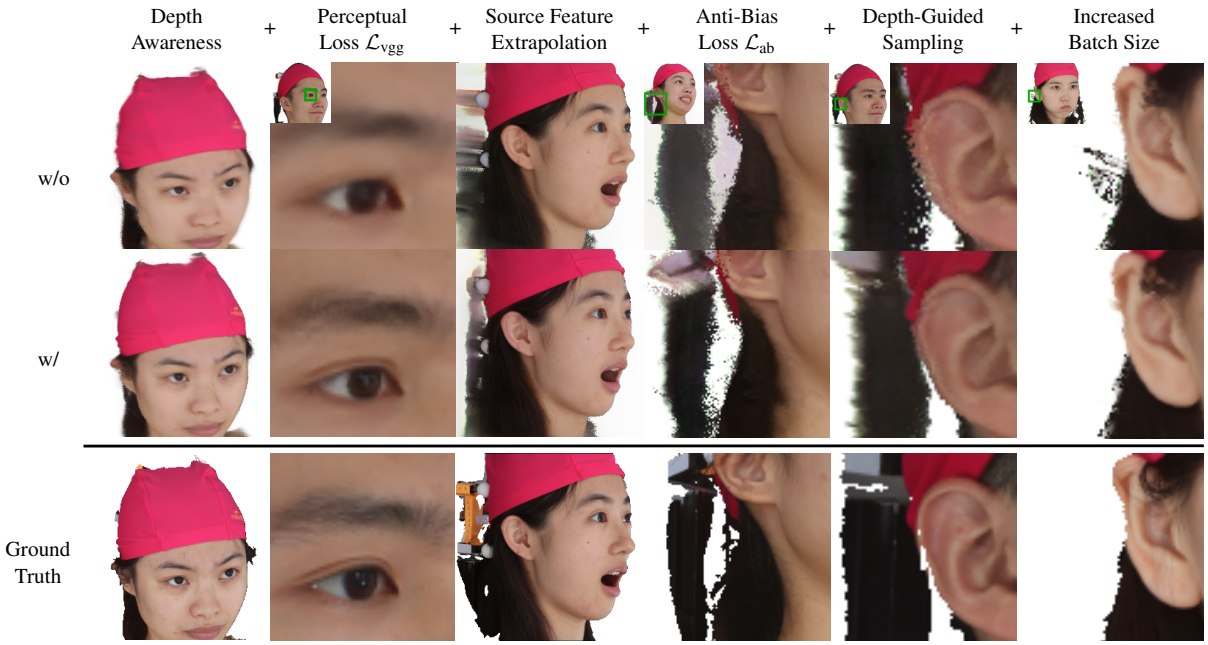

Figure 5. Qualitative ablation study. Starting with pixelNeRF [55] (first row, first column) as a baseline, we progressively add the components of our method and demonstrate their effects until we reach our final model (second row, last column).

trait photographs of 359 subjects under 20 different expressions captured with a rig of approximately 60 cameras, providing camera parameters and a 3D mesh that can be used to render the ground truth depth maps and segmentations for training. Four of the captured subjects consented to the public display of their data which determined our training and validation splits. We apply post-processing in terms of color calibration and unification of depth ranges, crop the images around the face region and rescale it to 256×256 px. We sample source view quadruples spanning horizontal and vertical angles of 45° and 30° respectively. Target views are sampled between the source views which results in $500k$ samples for training and $7k$ for validation.

Baselines We compare our approach against the state-of-the-art methods pixelNeRF [55], IBRNet [46], and KeypointNeRF [25] using the author's codebases. pixelNeRF [55] enables NeRFs to generalize across scenes by conditioning them on feature maps extracted from posed source views. IBRNet [46] adds a transformer architecture and estimates blend weights for the source images instead of regressing color directly. KeypointNeRF [25] adds triangulated landmarks to guide the geometry estimation. In the official implementation, KeypointNeRF only renders the intersection of the source view frustums which we changed to their union in order to be able to cope with partial observations in source views. Since off-the-shelf keypoint detectors struggle with strong head rotations, we provide KeypointNeRF with ground truth keypoints. We also evaluate a version of our method that does not include the perceptual loss and anti-bias term (see Section 4.4).

Method	LPIPS ↓	PSNR ↑	SSIM ↑	L1 ↓	L2 ↓
IBRNet [46]	0.159	22.7	0.89	0.025	0.006
pixelNeRF [55]	0.165	23.54	0.90	0.021	0.005
KeypointNeRF [25]	0.148	18.39	0.86	0.036	0.017
Ours w/o $\mathcal{L}_{\text{vgg}}$	0.137	**24.40**	**0.92**	**0.018**	**0.004**
Ours	**0.099**	22.42	0.91	0.020	0.007

Table 1. Quantitative comparisons on FaceScape show that our method has a significantly lower perceptual error in comparison to state-of-the-art methods while having on-par pixel-wise errors.

Metrics We quantitatively evaluate the performance of our model through the pixel-wise Manhattan and Euclidean distances (L1) and (L2), structural similarity (SSIM), learned perceptual image patch similarity (LPIPS), and the peak signal-to-noise ratio (PSNR).

5.1. Novel View Synthesis

Figure 4 displays a qualitative comparison between our method and the baselines on the FaceScape dataset [53]. IBRNet [46] generally performs well in regions that project onto large areas on the source images, e.g., wrinkles in the mouth region for extreme expressions are synthesized convincingly. However, especially the eye regions project onto small areas and IBRNet fails to produce plausible results. Regions that lie outside of the source views show significant artifacts. pixelNeRF [55] solves these cases better but tends to produce blurry results. KeypointNeRF [25] synthesizes high-frequency details very plausibly but shows artifacts for

Method	LPIPS ↓	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑
pixelNeRF [55]	0.16	0.021	**0.005**	23.54	0.90
+ Depth Awareness	0.15	**0.020**	**0.005**	**23.71**	**0.91**
+ Perc. Loss $\mathcal{L}_{\text{vgg}}$	0.11	0.032	0.008	21.66	0.89
+ Source Feature Extrapolation	0.11	0.029	0.008	21.96	0.90
+ Anti-bias Loss $\mathcal{L}_{\text{ab}}$	0.11	0.022	0.008	22.02	0.90
+ Depth-Guided Sampling	0.12	0.023	0.008	21.95	0.90
+ Increased Batch Size	**0.10**	**0.020**	0.007	22.42	**0.91**

Table 2. Quantitative ablation study on FaceScape [53].

regions outside of the source views and we also noticed severe deformations of the overall head shape. We attribute these deformations to the sparsity of the triangulated landmarks and their focus on the face area which leaves the remaining head regions without geometry information. The dense guidance by depth maps in our method effectively solves this artifact and similarly to pixelNeRF, regressing color values directly allows us to plausibly synthesize regions that lie outside of the source views. At the same time, even without a perceptual loss, our method synthesizes high-frequency details better than IBRNet and pixelNeRF. Adding a perceptual loss emphasizes high-frequency detail synthesis even more and yields results that qualitatively outperform all baselines even though understandably pixel-wise scores slightly worsen (see Table 1).

5.2. Ablation Study

Since our approach is based on pixelNeRF [55], we perform an additive ablation study to evaluate the contributions of our changes in which we progressively add our novel components and discuss their effects on the model performance. Figure 5 visualizes the qualitative effects by progressively introducing one new component per column. Table 2 provides the quantitative evaluation. First, we add depth awareness to pixelNeRF (Section 4.1) which improves the overall synthesis quality and all scores. Introducing the perceptual loss $\mathcal{L}_{\text{vgg}}$ adds more high-frequency details but pixel-wise scores degrade slightly. Source feature extrapolation (Section 4.2) counters smearing artifacts that occur in regions that are invisible in the source views. Color shifts that appeared after adding $\mathcal{L}_{\text{vgg}}$ can be eliminated with the anti-bias loss $\mathcal{L}_{\text{ab}}$ (Section 4.4). Depth-guided sampling (Section 4.3) increases the sampling density around the head surface and especially improves the synthesis quality of thin surfaces like ears. As a side effect, it also allows us to reduce the number of samples per ray and increase the batch size from 1 to 4 without changing GPU memory requirements. This stabilizes training such that minor artifacts vanish and consistently improves all metrics.

5.3. Novel View Synthesis on General Objects

Since depth maps can be estimated for arbitrary scenes, we can evaluate our model on general objects in the DTU dataset. The DTU dataset [18] is a large-scale multi-view data set of general scenes under controlled conditions. It

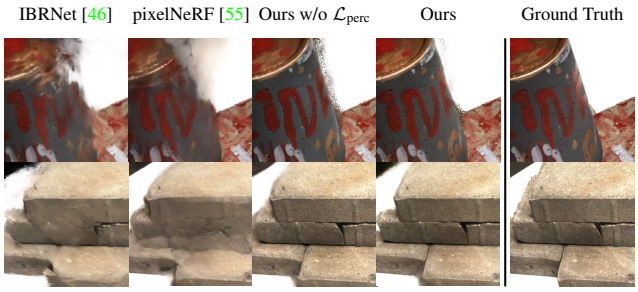

Figure 6. Qualitative comparison on DTU [18] which consists of a variety of objects like a bucket (1rst row) or bricks (2nd row).

contains 124 scenes with general objects captured from 49 to 64 positions including ground truth depth. We follow the training/validation split convention proposed in [55], adopt the preprocessing steps from [8], and apply additional 2-fold downsampling to obtain images with resolution 256×320 px. Similar to the training on FaceScape, we sample source view quadruples spanning horizontal and vertical angles of $\approx 50°$ and $\approx 35°$ respectively which results in $30k$ samples for training and $5k$ for validation. In Figure 6, we show the comparison of our method to the baselines. Note that KeypointNeRF is class-specific and can not be applied to general objects. The results demonstrate that our method outperforms all baselines by a significant margin, see Table 3 for the quantitative evaluation.

6. Discussion

DINER excels at synthesizing photo-realistic 3D scenes from few input images. Still, some challenges remain before it may be used for real-world applications such as immersive video conferencing. Similar to most NeRF-based methods, our rendering speed is slow. Despite improved sampling efficiency through depth guidance, the synthesis of a 256^2 px image still takes two seconds. While real-time-capable NeRF methods exist [13, 28, 54], none of them generalizes across scenes yet. Applying our method to dynamic scenes and confirming temporal stability would be highly interesting, yet at the time of developing DINER, no public dataset provides dynamic captures for a high number of subjects. As our method relies on a depth prediction network, we are bound by its accuracy. However, it could be replaced by depth measurements from Kinect-like sensors.

7. Conclusion

We presented depth-aware image-based neural radiance fields (DINER) which synthesize photo-realistic 3D scenes given only four input images. To capture scenes more completely and with high visual quality, we assume to have input images with a high disparity. We leverage a state-of-the-art depth estimator to guide the implicit geometry estimation and to improve sampling efficiency. In addition,

we propose a technique to extrapolate features from the input images. Our experiments show that DINER outperforms the state of the art both qualitatively and quantitatively. DINER's ability to reconstruct both human heads as well as general objects with high quality is vital for real-world applications like immersive video conferencing with holographic displays.

8. Acknowledgements

This project has received funding from the Max Planck ETH Center for Learning Systems (CLS). Further, we would like to thank Marcel C. Bühler, Philip-William Grassal, Berna Kabadayi, Jalees Nehvi, Balamurugan Thambiraja, and Wojciech Zielonka for their valuable feedback.

References

[1] 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. Mar. 2021. 2

[2] Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[3] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shunsuke Saito, Stephen Lombardi, Shih en Wei, Danielle Belko, Shoou i Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *SIGGRAPH*, 2022. 1, 3

[4] 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. Eg3d: Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 2

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 2

[6] 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, 2021. 1, 2

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 4

[8] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2, 4, 8

[9] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars, 2022. 2, 3

[10] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. volume 40, 2021. 3

[11] Guy Gafni, Justus Thies, Michael Zollhoefer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. Dec. 2020. 1, 2, 3

[12] Stephan J. Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. Voltemorph: Real-time, controllable and generalisable animation of volumetric representations, 2022. 3

[13] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps, 2021. 2, 8

[14] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. *CoRR*, abs/2112.01554, 2021. 1, 3

[15] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5151–5160, June 2021. 1

[16] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model, 2021. 3

[17] Tao Hu, Shu Liui, Lun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf – efficient neural radiance fields. *CVPR*, 2022. 2

[18] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Dtu: Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 2, 8, 12, 13

[19] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzciński, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields, 2021. 2

[20] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars, 2022. 3

[21] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018. 3

[22] Haotong Lin, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields with learned depth-guided sampling, 2021. 2

[23] 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. 3

[24] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mix-

ture of volumetric primitives for efficient neural rendering, 2021. 1, 3

[25] 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. 1, 2, 3, 6, 7, 13

[26] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16190–16199, June 2022. 2

[27] 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. 2, 3, 4, 12, 14

[28] Thomas Mueller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding, 2022. 2, 8

[29] 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. 2

[30] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[31] 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. 2

[32] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1

[33] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. jun 2021. 2

[34] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giró-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 (ICCV)*, pages 5620–5629, October 2021. 3

[35] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. *arXiv preprint arXiv:2111.14643*, 2021. 2, 4

[36] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2, 4

[37] Radu Alexandru Rosu, Shunsuke Saito, Ziyan Wang, Chenglei Wu, Sven Behnke, and Giljoo Nam. Neural strands: Learning hair geometry and appearance from multi-view images. *ECCV*, 2022. 1, 3

[38] Mehdi S. M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lucic, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene Representation Transformer: Geometry-Free Novel View Synthesis Through Set-Latent Scene Representations. *CVPR*, 2022. 1

[39] Aliaksandr Siarohin. First order motion model for image animation. *Advances in Neural Information Processing Systems*, 32:7137–7147, jun 2019. 2, 3

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 5

[41] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Generalizable patch-based neural rendering. In *European Conference on Computer Vision*. Springer, 2022. 1

[42] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022. 2

[43] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in neural rendering. 2022. 1, 3

[44] J. Thies, A. Tewari, O. Fried, V. Sitzmann, S. Lombardi, K. Sunkavalli, R. Martin-Brualla, T. Simon, J. Saragih, M. Nießner, R. Pandey, S. Fanello, G. Wetzstein, J.-Y. Zhu, C. Theobalt, M. Agrawala, E. Shechtman, D. B Goldman, and M. Zollhöfer. State of the art on neural rendering. *EG*, 2020. 1

[45] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[46] 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. 1, 2, 6, 7, 8, 12, 13

[47] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *European Conference on Computer Vision*, 2022. 1

[48] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[49] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5704–5713, June 2021. 3

[50] Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Michael Zollhoefer, Jessica Hodgins, and Christoph Lassner. Hvh: Learning a hybrid neural volumetric representation for dynamic hair performance capture, 2021. 1, 3

[51] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. *arXiv preprint arXiv:2201.04127*, 2022. 1

[52] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. Jan. 2022. 2

[53] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 5, 7, 8, 13, 15

[54] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 8

[55] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8, 12, 13

[56] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference of Computer vision (ECCV)*, pages 524–540, August 2020. 2, 3

[57] Yufeng Zheng, Victoria Fernández Abrevaya, Xu Chen, Marcel C. Buehler, Michael J. Black, and Otmar Hilliges. I m avatar: Implicit morphable head avatars from videos. Dec. 2021. 1, 3

[58] Michael Zollhöfer, Justus Thies, Darek Bradley, Pablo Garrido, Thabo Beeler, Patrick Péerez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. 2018. 3

DINER: Depth-aware Image-based NEural Radiance fields
– Supplemental Document –

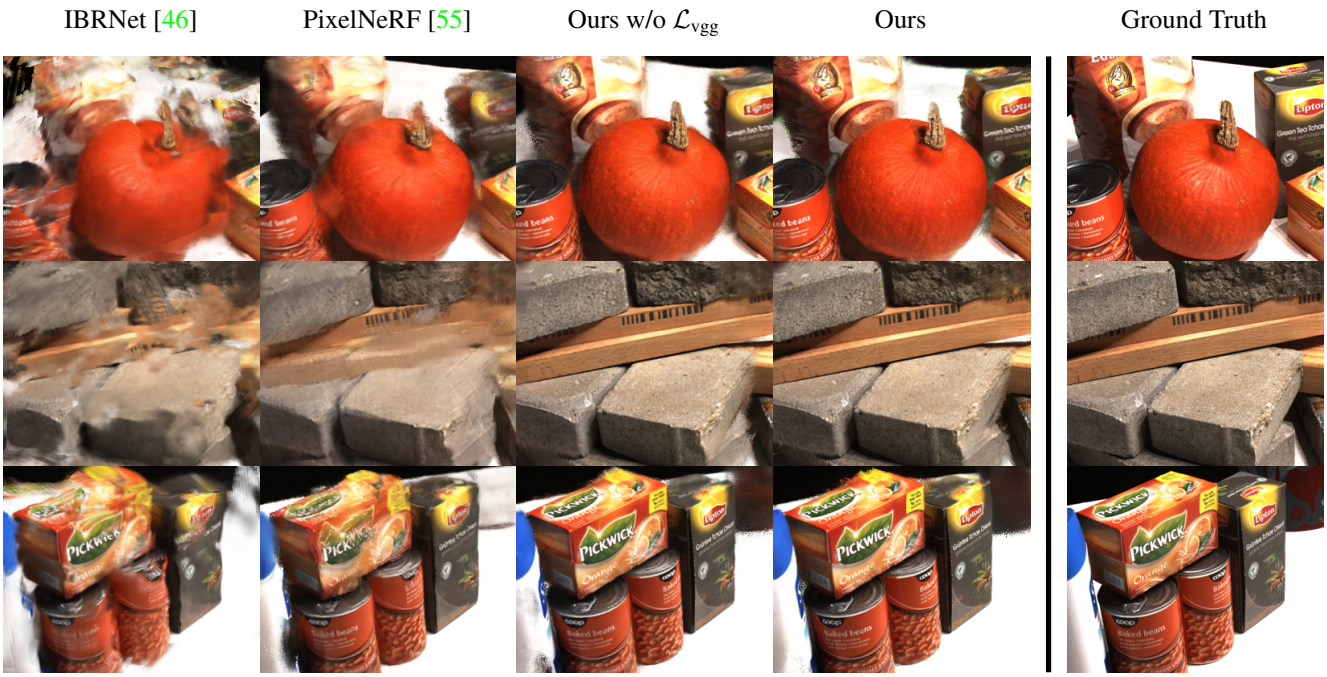

Figure 7. Qualitative comparison on general objects of the DTU dataset [18]. Our depth-aware image-based neural radiance field shows significantly higher image quality with fewer distortions and blurring artifacts.

Abstract

In this supplemental document, we detail the architecture of our method DINER *(see Appendix A), provide a quantitative comparison to state-of-the-art models on novel view synthesis for general objects in the DTU dataset (see Appendix B), evaluate the influence of the depth estimator's accuracy on the synthesis quality (see Appendix C), and conduct further experiments concerning depth-guided sampling (see Appendix D). We conclude this document with a discussion of ethical implications of our work (see Appendix E).*

A. Architecture Details

We adopt the model architecture of pixelNeRF [55] and kindly refer to their supplemental material for further details about the image encoder and the NeRF network. Our newly introduced components require two adaptations, namely when we introduce depth awareness we change the dimensionality of the feature vector that conditions the MLP, and the source feature extrapolation requires us to change the input channel size of the image encoder. Both adaptations will be detailed in the following paragraphs.

Depth Awareness To guide the scene reconstruction, we also condition the NeRF on the positionally encoded distance between the z-coordinate of the sampling point in camera coordinates and the projected depth value. We employ the same positional encoding as in the original NeRF [27] and use 6 frequency channels with a base frequency of $1\frac{1}{\text{meter}}$. The resulting 13-dimensional vector is concatenated with the 512-dimensional feature vector sampled from the feature maps and then used to condition the NeRF MLP. The input layer weight dimensions of the MLP are adjusted accordingly.

Source Feature Extrapolation We use a combination of image padding and positional encoding to enable the image encoder to extrapolate the feature maps. The images are padded by 64 px by repeating the border values. The positional encoding ranges over 4 exponentially increasing

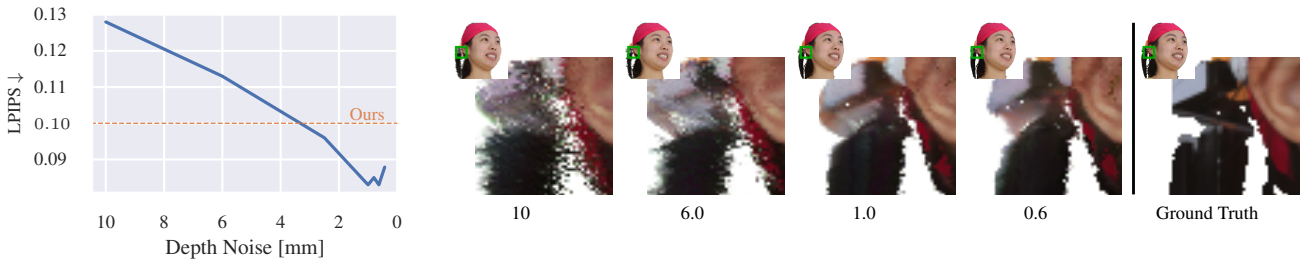

Figure 8. Model performance under noisy depth signals. The synthesis quality improves with increasing depth accuracy up to a standard deviation of 1mm. Depth information with even higher accuracy does not yield further improvements in terms of synthesis quality.

frequencies starting with 0.5 and is applied to the pixel's uv coordinates which are normalized to $[-1, +1]$. The resulting positional encoding map has a channel size of 18. Note that the positional encoding is set to 0 for all pixels that do not belong to the padded region. Adding positional encodings to the source image before applying the image encoder means that the inputs to the image encoder no longer have 3 channels. Since we employ a pretrained network, we have to add randomly initialized weights to its first layer. Note that because the positional encoding maps are set to zero in unpadded regions, here the added weights do not have an effect on the predictions of the pretrained network.

Depth-Guided Sampling For depth-guided sampling, we use 1000 candidate samples per ray from which we shortlist 25 samples and add 15 samples during Gaussian boosting. This sums up to 40 samples in total which contribute to the final ray color. The normal maps that we require for point cloud backface culling are obtained by calculating the central difference on the depth maps via convolutional kernels with size 3. Foreground-background edges are filtered out.

Objective Function The objective function for training DINER consists of 3 terms: a pixel-wise l_1 distance $\mathcal{L}_{l_1}$, a perceptual loss $\mathcal{L}_{\text{vgg}}$, and the anti-bias term $\mathcal{L}_{\text{ab}}$. The according weights are

$$w_{l_1} = 1.0$$
$$w_{\text{vgg}} = 0.1$$
$$w_{\text{ab}} = 5.0 \text{ (1.0 for DTU)}.$$

All terms are evaluated on patches of 64×64 px unless noted otherwise. $\mathcal{L}_{\text{ab}}$ downsamples the patches to 8×8 px through average pooling before evaluating the l_1 distance. The perceptual loss was adopted from [25].

B. Quantitative Comparison on DTU

We presented a qualitative comparison for novel view synthesis of general objects in the DTU dataset [18] in the main paper and in Figure 7. The quantitative evaluation is provided in Table 3. Please note that KeypointNeRF [25] cannot be applied to general objects since keypoints cannot be generalized to arbitrary objects. Our method outperforms all baseline methods by a significant margin. The improvements are even more noticeable than for the FaceScape dataset [53] for which we presented the quantitative results in the main paper. We found that while previous methods are able to learn a coarse geometry prior when applied to heads only, i.e. when trained and evaluated on FaceScape, they fail to do so for general scenes. As a consequence, exploiting depth information to guide the synthesis of general scenes is even more beneficial. On the other hand, we found that adding a perceptual loss does not increase the synthesis quality as much as for FaceScape.

Method	LPIPS ↓	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑
IBRNet [46]	0.40	0.066	0.017	19.94	0.65
pixelNeRF [55]	0.38	0.055	0.011	20.96	0.67
KeypointNeRF [25]	–	–	–	–	–
Ours w/o $\mathcal{L}_{\text{vgg}}$	0.27	**0.037**	**0.006**	**24.14**	**0.82**
Ours	**0.23**	0.039	0.007	23.44	0.81

Table 3. Quantitative comparison on DTU [18].

C. Influence of Depth Accuracy

Since our method relies on predicted depth maps which are subject to inaccuracies, we investigate how depth accuracy reflects on the synthesis quality. To this end, we perform a set of experiments where we train our model on the ground truth depth perturbed by Gaussian noise with varying standard deviations. Figure 8 displays the quantitative and qualitative findings. We observe that higher depth accuracy also improves the synthesis quality up until a standard deviation of 1mm. More accurate depth information does not improve synthesis quality further. We conclude that a better depth estimation network could yield an additional boost to our model's performance.

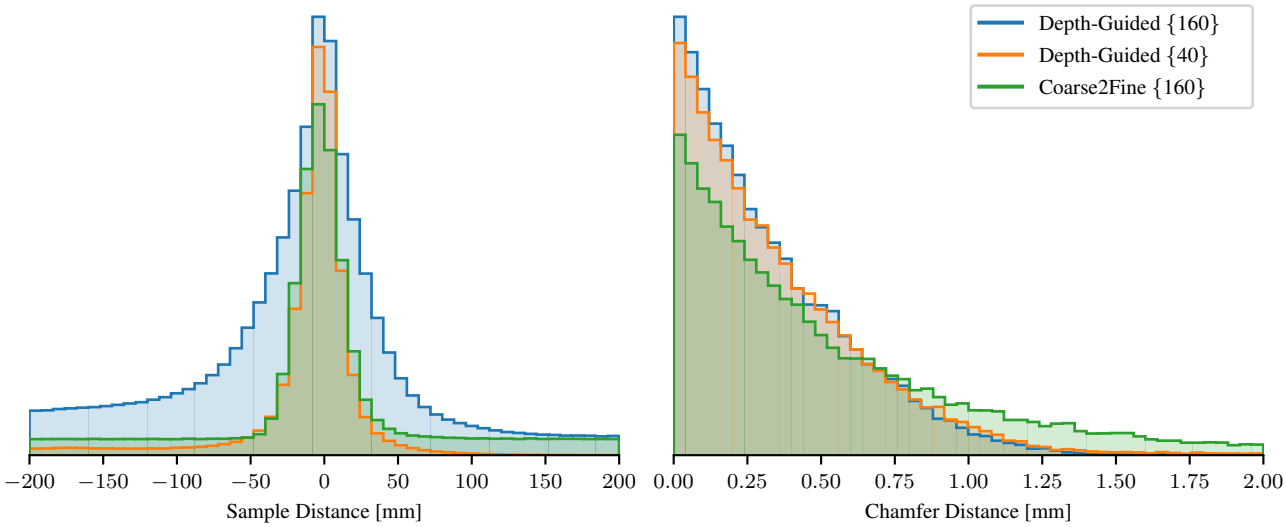

Figure 9. Distances between sampled points and ground truth surface for depth-guided sampling and standard coarse-to-fine-based sampling as in the original NeRF paper [27]. Curly braces indicate the number of samples per ray. Left: distances between sampling points and ground truth surface. Right: distances between ground truth surface and closest sampling point (*Chamfer distance*). Depth-guided sampling effectively focuses the sampling on the ground truth surface and places samples closer to it.

Sampling Strategy	Median Chamfer Dist.	Maximum Chamfer Dist.
Coarse-to-Fine {160}	0.39 mm	6.7 mm
Coarse-to-Fine {40}	6.4 mm	55.6 mm
Depth-Guided {160}	0.26 mm	1.6 mm
Depth-Guided {40}	0.28 mm	6.7 mm

Table 4. Distances between the ground truth surface and the closest sampling points (*Chamfer distance*) for different sampling strategies. Curly brackets indicate the number of samples per ray. Depth-guided sampling places samples closer to the ground truth surface and focuses on these areas even if only few samples are drawn.

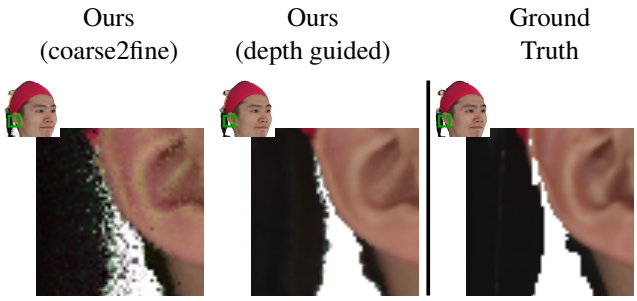

Figure 10. Qualitative comparison of sampling strategies. Both models sample only 40 points per ray and were trained with batch size 4. Depth guidance improves sampling efficiency and solves artifacts around thin surfaces.

D. Depth-Guided Sampling

In this section, we analyze how depth guidance improves sampling efficiency. More specifically, we measure how close the sampled points lie around the ground truth surface. For this, we consider two quantities: the distances between sampled points and the ground truth surface, and the distances between the ground truth surface and closest sampling points, i.e., the Chamfer distance. Figure 9 visualizes both distributions in comparison to the standard coarse-to-fine sampling strategy as introduced in the original NeRF paper [27]. In Figure 9(left), we observe that coarse-to-fine sampling places a comparably small number of samples close the to ground truth surface. This is because first, a partition of the samples must be used to query the coarse MLP to find regions of interest; second, even a part of the remaining samples is used to uniformly query the space which leads to long, non-vanishing tails in the distance distributions. As a consequence of the low sample density around the ground truth surface, we observe fewer surface points with small Chamfer distances in Figure 9(right) and a comparatively high median Chamfer distance in Table 4. In contrast, depth-guided sampling with the same number of points per ray places more samples closer to the ground truth surface (see Figure 9), which reduces the median Chamfer distance by 33% and the maximum Chamfer distance by a factor of 4 (see Table 4). Note that depth-guided sampling does not require querying a coarse MLP and, therefore, more samples contribute directly to the final output color. Even when we reduce the number of samples by a factor of 4, Figure 9(left) shows that depth-guided sampling focuses on areas close to the ground truth surfaces and predominantly minimizes the tails of the distance distribution, i.e., drops samples that lie far away from the surface. As a consequence, compared to standard coarse-to-

fine sampling with 4 times more samples, we observe a significantly improved median Chamfer distance (see Table 4). In contrast, when cutting the number of samples per ray for standard coarse-to-fine sampling, we observe significantly degraded Chamfer distances. Figure 10 demonstrates that this results in severe artifacts around thin surfaces during novel view synthesis. We conclude that only depth-guided sampling allows us to cut the number of sampled points per ray by a factor of 4 without introducing artifacts. This in turn allows us to increase the batch size during training from 1 to 4 without changing hardware requirements which we found to improve model performance.

E. Ethical Considerations

Our method reconstructs a volumetric representation of a subject or general objects from sparse color camera inputs. Since this volumetric representation does only allow for novel view synthesis, there is no immediate risk of misuse, such as deep fakes. As no personalized avatar is reconstructed, a potential immersive telepresence application does not need to store person-specific information. We train the method on FaceScape [53] which is not a balanced face dataset and is biased towards the local population. There is no explicit usage of this bias in the proposed method, thus we expect that our method can also be trained on a diverse dataset of humans which was not available at the development time.

The human data used in this study is based on the FaceScape dataset with the consent of the subjects to be used for research. Four subjects agreed to be displayed in publications and presentations; these subjects are the test set.