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

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

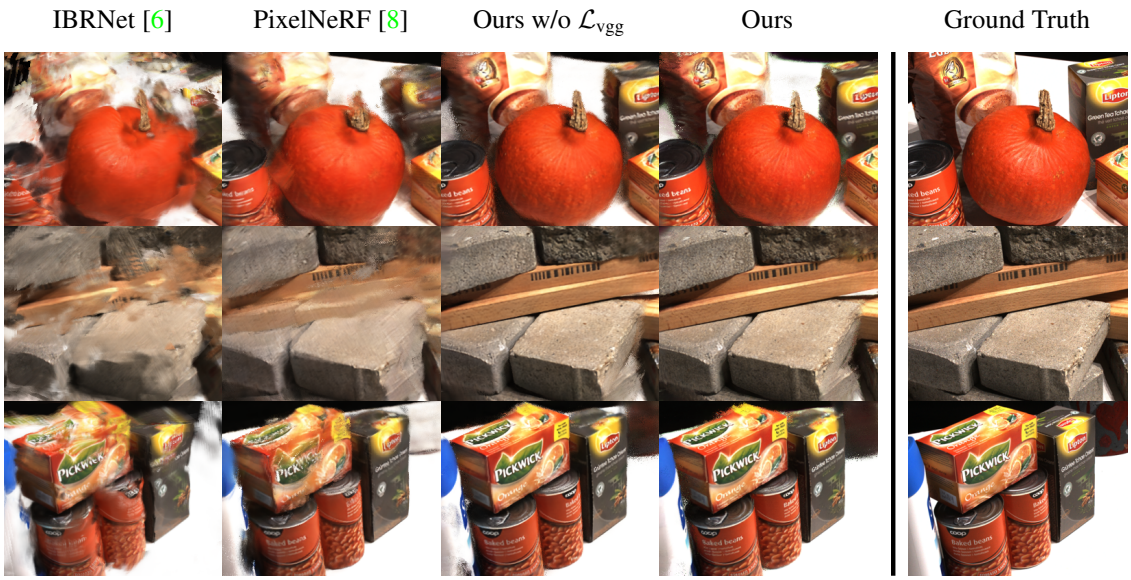

Figure 1. Qualitative comparison on general objects of the DTU dataset [2]. Our depth-aware image-based neural radiance field shows significantly higher image quality with fewer distortions and blurring artifacts.

Abstract

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

A. Architecture Details

We adopt the model architecture of pixelNeRF [8] 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.

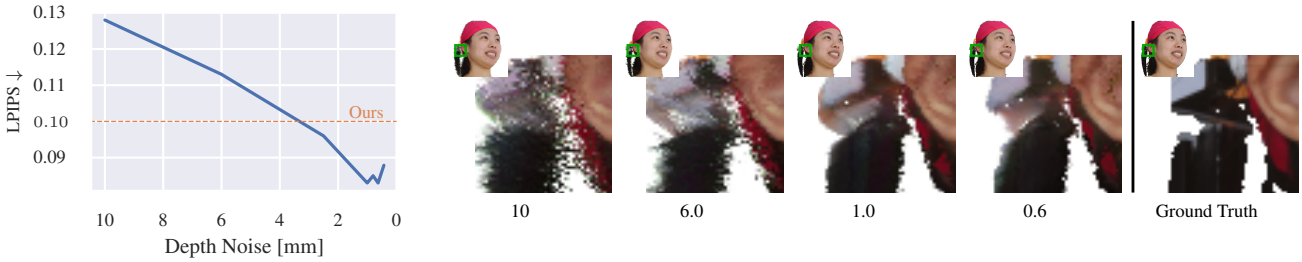

Figure 2. Model performance under noisy depth signals. The synthesis quality improves with increasing depth accuracy up to a standard deviation of 1mm. Depth information with even higher accuracy does not yield further improvements in terms of synthesis quality.

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 [5] 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 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

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

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 [4].

B. Further Comparisons on DTU

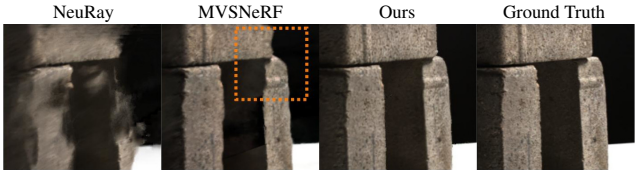

Figure 3. Qualitative comparison to NeuRay [3] and MVS-NeRF [1] on DTU [2].

Method	LPIPS ↓	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑
NeuRay	0.41	0.069	0.017	19.50	0.65
MVSNeRF	0.35	0.059	0.013	20.45	0.67
IBRNet [6]	0.40	0.066	0.017	19.94	0.65
pixelNeRF [8]	0.38	0.055	0.011	20.96	0.67
KeypointNeRF [4]	–	–	–	–	–
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 1. Quantitative comparison on DTU [2].

We presented a qualitative comparison for novel view synthesis of general objects in the DTU dataset [2] in the main paper and in Figure 1. The quantitative evaluation is provided in Table 1. Please note that KeypointNeRF [4] cannot be applied to general objects since keypoints cannot be generalized to arbitrary objects and that we added two further baseline methods NeuRay [3] and MVS-NeRF [1]. The qualitative comparison between DINER and

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 2. 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.

NeuRay and MVSNeRF can be found in Fig. 3. Our method outperforms all baseline methods by a significant margin. The improvements are even more noticeable than for the FaceScape dataset [7] 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.

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 2 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.

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 4 visualizes both distributions in comparison to the standard coarse-to-fine sampling strategy as introduced in the original NeRF paper [5]. In Figure 4(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 4(right) and a comparatively high median Chamfer distance in Table 2. In contrast, depth-guided sampling with the same number of points per ray places more samples closer to the ground truth surface (see Figure 4), which reduces the median Chamfer distance by 33% and the maximum Chamfer distance by a factor of 4 (see Table 2). 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 4(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 2). In contrast, when cutting the number of samples per ray for standard coarse-to-fine sampling, we observe significantly degraded Chamfer distances. Figure 5 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 [7] which is not a balanced face dataset and is biased towards the local population. However, in the main paper, we show that DINER generalizes well to subjects of unseen ethnicities and therefore rules out discrimination against underrepresented minorities.

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.

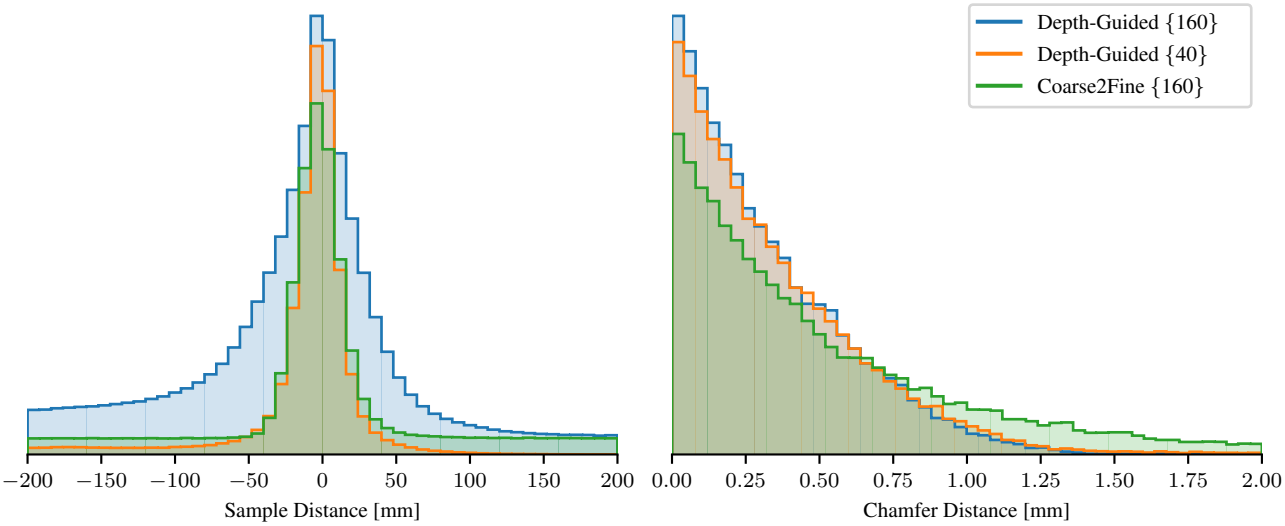

Figure 4. Distances between sampled points and ground truth surface for depth-guided sampling and standard coarse-to-fine-based sampling as in the original NeRF paper [5]. Curly braces indicate the number of samples per ray. Left: distances between sampling points and ground truth surface. Right: distances 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.

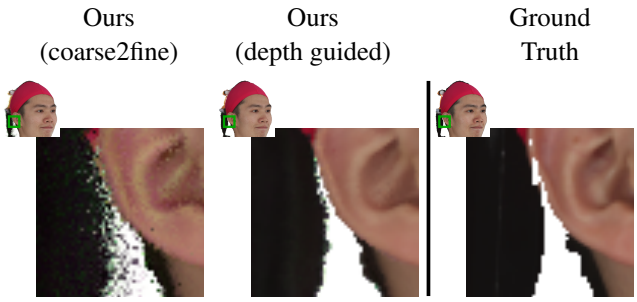

Figure 5. Qualitative comparison of sampling strategies. Both models sample only 40 points per ray and were trained with batch size 4. Depth guidance improves sampling efficiency and solves artifacts around thin surfaces.

References

[1] Anpei Chen et al. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo, 2021. 2

[2] 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. 1, 2

[3] Yuan Liu et al. Neural rays for occlusion-aware image-based rendering. In *CVPR*, 2022. 2

[4] 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. 2

[5] 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

[6] 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

[7] 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. 3

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