Depth-guided NeRF Training via Earth Mover's Distance

Anita Rau, Josiah Aklilu, F. Christopher Holsinger, and Serena Yeung-Levy

Stanford University
{arau, josaklil, holsinger, syyeung}@stanford.edu
https://anitarau.github.io/emd-nerf.github.io/

Abstract. Neural Radiance Fields (NeRFs) are trained to minimize the rendering loss of predicted viewpoints. However, the photometric loss often does not provide enough information to disambiguate between different possible geometries yielding the same image. Previous work has thus incorporated depth supervision during NeRF training, leveraging dense predictions from pre-trained depth networks as pseudo-ground truth. While these depth priors are assumed to be perfect once filtered for noise, in practice, their accuracy is more challenging to capture. This work proposes a novel approach to uncertainty in depth priors for NeRF supervision. Instead of using custom-trained depth or uncertainty priors, we use off-the-shelf pretrained diffusion models to predict depth and capture uncertainty during the denoising process. Because we know that depth priors are prone to errors, we propose to supervise the ray termination distance distribution with Earth Mover's Distance instead of enforcing the rendered depth to replicate the depth prior exactly through L_2-loss. Our depth-guided NeRF outperforms all baselines on standard depth metrics by a large margin while maintaining performance on photometric measures.

Keywords: Neural radiance fields · Depth prediction · Monocular depth priors · Earth Mover's Distance

1 Introduction

Neural Radiance Fields (NeRFs) [15] have demonstrated an impressive ability to render novel views of a known scene. Especially in object-centric and well-sampled scenes, NeRFs can generate photometrically and geometrically consistent images from previously unseen view points. However, in camera-centric and sparse view scenarios, neural radiance fields have yet to show the same fidelity. Additionally, while producing renderings of high quality, some NeRFs fail to capture the underlying geometry of a scene accurately, which is essential for applications in robotics or augmented reality [1, 3, 27, 36]. Some reasons include a smaller overlap between images, occlusions between views, and photometric inconsistencies due to the camera's auto exposure.

To improve the robustness of NeRFs in complex indoor settings, previous work [7, 22, 25, 26] has incorporated depth supervision during training. The idea

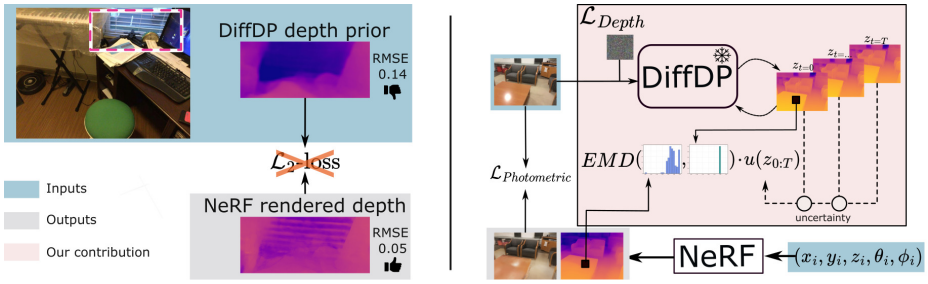


Fig. 1: Left: Predicted monocular depth priors are not perfect and false interpretations of a scene's geometry are unavoidable. Blindly forcing a NeRF to replicate such priors (e.g. through L_2-loss) leads to high geometric losses. Right: Overview of our method. We use depth priors to guide NeRF training via Earth Mover's Distance (EMD).

is that a better understanding of the underlying scene geometry will enhance image rendering. However, dense depth ground truths are rarely available.

Fortunately, a line of work has shown that depth estimation networks have become accurate at predicting the layout of indoor environments [2,11,17,20,21] and they can be used as pseudo ground truth for NeRF training. But when used ineffectively, depth supervision can fail to improve the NeRF's depth renderings, suggesting a deficient understanding of scene geometry. One reason is that, although depth prediction with neural networks has improved dramatically, there is still ambiguity owed to the ill-defined nature of estimating depth from one view, as many geometric layouts can lead to the same image. Partial occlusion of objects, shading, and reflections further challenge monocular depth prediction, leading to misinterpreted geometries.

Several methods have addressed the described ambiguity in monocular depth. Some works assume that, once a depth prior is accepted, all pixels within an image are strong pseudo ground truths. [26] proposed to incorporate expensive, custom-trained multi-modally distributed depth hypotheses of which one is picked for supervision at each training step. However, we observe that generating additional depth predictions comes at the price of generating harmful noise. [25] treats patches individually, but assumes that all pixels in a patch are either a perfect ground truth or not useful at all, when in practice all pixels have varying levels of uncertainty. Other methods do consider uncertainty at the pixel-level, but treat depth priors as normally distributed around the true depth, which is over-simplified and empirically does not hold [7,22]. To our knowledge, none of the previous methods have captured notions of uncertainty in depth priors and leveraged them for NeRF training effectively.

We argue that although monocular depth predictions can be highly inaccurate, some are useful—so it is crucial to avoid a strict adherence to the depth prior in NeRF's predictions. Depth priors should be a *suggestion*. We propose to supervise the ray termination distance of a NeRF with the Earth Mover's Distance (EMD) as illustrated in Figure 1. This invites the NeRF to sample

ray termination distances close to the depth prior, without directly competing for the final weighting of ray termination distances with the photometric loss. Unlike previous works, we avoid L_2-losses that will, inevitably, enforce incorrect depth as well, such as the missed window blinds in Figure 1. Instead, EMD allows us to maintain useful information about the distribution of the ray termination distances, while avoiding restrictive assumptions such as normally distributed errors, unimodality, continuity, or non-zero probabilities such as KL-divergence.

Our work employs out-of-the-box pre-trained generative diffusion models as depth priors and leverages the denoising process for additional uncertainty estimation. This provides accurate depth priors and uncertainties for free. These uncertainty maps tell us when the depth model is unsure, and when we should rely more on the RGB loss. We then introduce an approach to weighing our novel depth loss and the photometric loss that is inspired by Focal Loss [14].

In summary, we propose a new way to think about uncertainty in depth supervised NeRF. We provide novel techniques to incorporate pixel-wise uncertainty, without imposing restrictive assumption on the nature or distribution of the uncertainty, the depth prior, or NeRF ray termination distances. We outperform all baselines on all depth metrics by at least 11% on ScanNet [5], and outperform the most closely related baseline by up to 54% on the relative error. The results speak to our method's ability to understand the underlying geometry of a scene rather than just rendering accurate looking images based upon an incorrect understanding of the 3D world.

2 Related Work

Monocular depth estimation: Depth estimation from a single view is inherently ambiguous but many works have achieved impressive depth accuracy [2, 11, 13, 23, 30, 31, 34]. [2] describes a method for depth estimation that generalizes well to multi-domain data by bridging the gap between relative and metric depth estimation. [31] uses a large-scale dataset and canonical camera transformations on input views to resolve the same metric ambiguity in zero-shot. [17, 29, 30] demonstrate how rich representations from self-supervised pretraining enables strong performance in dense visual perception. Most recently, [11, 23] leverage image-conditioned denoising diffusion in the depth estimation pipeline. In our work, we capitalize on the robust uncertainty estimates readily available in generative approaches to depth prediction. We use the monocular depth estimation network DiffDP [1] [11] to provide depth pseudo ground truths for NeRF.

NeRF with sparse views: Neural radiance fields [15] have become a popular choice for 3D scene representations due to their ability to render accurate novel views. NeRF comprises of a neural network that takes as input coordinates and camera parameters and outputs color and density. However, the original NeRF relies crucially on many inputs views from the scene of interest in order to faithfully reconstruct the scene geometry. Additionally, it is known that naively

[1] Both [11] and [22] are called DDP, so we refer to them as DiffDP and DDPrior.

training NeRF tends to oversample empty space in the scene volume, leading to artifacts and inaccurate depth prediction. In this work, we consider the setting of camera-centric indoor environments with a few dozen input images. As the overlap between images is low, additional regularization is crucial.

Thus, recent works [7, 9, 10, 16, 22, 26, 33] have enhanced NeRF when only sparse views are available. Many of these rely on data-driven priors or regularization to guide NeRF optimization. [33] equips NeRF with per-view image features from a CNN encoder that is trained on large multi-view datasets, enabling scene reconstruction from as little as one view in a single forward pass. [10] supervises NeRF by enforcing consistency amongst CLIP [19] representations of arbitrary views. [16] samples unobserved views and regularizes the geometry and appearance of rendered patches from these views.

NeRF with depth supervision: Amongst these, an interesting line of work explores supervising NeRF construction with depth [7,9,12,22,25,26,28]. Notably, few of these works explicitly identify the inherent uncertainty in depth predictions. [7] supervises NeRF ray sampling with COLMAP-derived [24] depth and models noise with a Gaussian centered at the depth predictions. [22] train a custom network in-domain to predict uncertain areas. However, the model overfits to the training case, where high depth errors occur at the edges of objects. [9, 18] focus on object-centric applications with large overlap between train images. [18] employs predicted depth by using the difference between sample location and estimated depth as conditioning for the NeRF and for depth-guided ray sampling. [9] supervises NeRF optimization by relaxing hard constraints in depth supervision to softer, more robust local depth ranking constraints. Although this approach acknowledges potential noise in pretrained depth network estimates, it does not directly incorporate uncertainty to guide depth supervision for NeRF. SCADE [26] leverages 20 ambiguity-aware depth proposals generated by a costly out-of-domain prior network and a space-carving loss with mode-seeking behavior to supervise ray termination distance in NeRF. We found that while some depth hypotheses can be accurate, many of these proposals may also be poor predictors of scene geometry. This finding challenges the utility of having multimodal depth proposals for resolving ambiguity (see Figure 8), when NeRF naturally captures multimodal ray termination distributions. DäRF [25] is most closely related to our work and, like our work, proposes to use a standard pretrained depth estimation network to supervise NeRF training. Yet, DäRF does not consider pixel-wise uncertainty and supervises the NeRF-rendered depth directly instead of guiding the ray termination distance distribution.

3 Method

An overview of our method is shown in Figure 2. In this section, we describe each of the three components of our pipeline (depth prior, EMD-guided ray termination sampling, and photometric-geometric loss balancing) in turn.

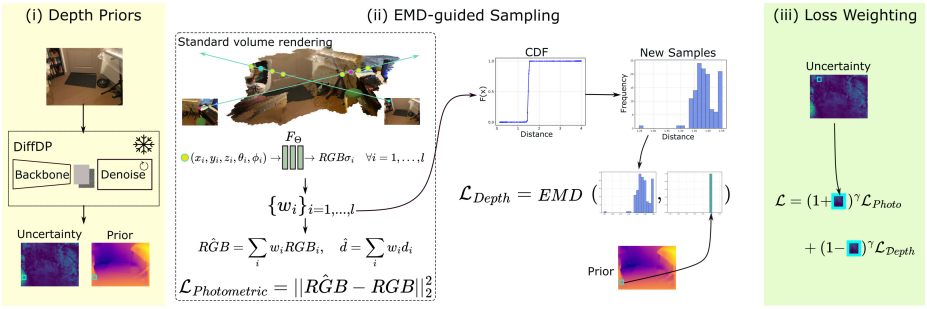


Fig. 2: A detailed schematic of our depth-guided NeRF optimization. (i) A pretrained diffusion model for depth prediction, DiffDP [11], provides depth priors. Measuring the progression of depth predictions throughout the denoising process provides uncertainty maps. (ii) Given inputs poses (x,y,z,θ,ϕ) a network F outputs RGB value and density. From the outputs, we derive weights w that, when normalized, serve as a piece-wise-constant probability density function. We can then construct a cumulative distribution function (CDF) from which we sample new ray termination distances. We supervise these samples with the Earth Mover's Distance (EMD) to the depth prior. (iii) Finally, we weigh the photometric and depth losses according to the DiffDP-derived uncertainty.

3.1 Background

The core of our work is a standard neural radiance field (NeRF) [15]. Given a collection of images $\{\mathbf{I}_i\}_{i=1,\dots,N}$ and their corresponding SfM camera poses, we aim to train a network to render images from novel, unseen views. Similar to existing works, we encode a NeRF as multilayer perceptron $F_\Theta : (x_i, y_i, z_i, \theta_i, \phi_i) \rightarrow (\mathbf{c_i}, \sigma_i)$ that takes camera position and viewing direction as input and outputs color $\mathbf{c}$ and volume density σ. To render depth from a given camera view, we cast a ray $\mathbf{r}$ through the origin $\mathbf{o}$ of the camera and a pixel projected into world space. We sample termination distances from the ray within a predefined interval $(\text{near}, \text{far})$ and pass them through a neural radiance field network that predicts their weights w_i as probability that the ray can traverse space without obstructions until reaching the hypothesis in question: $w_i = T_i(1 - \exp(-\sigma_i\delta_i))$, where δ_i is the distance between samples, and $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j\delta_j)$. The final predicted color $\hat{C}$ and depth $\hat{d}$ is then the weighted average of all l ray termination distances along $\mathbf{r}$

$$\hat{C}(\mathbf{r}) = \sum_l w_l \mathbf{c}_l \quad \text{and} \quad \hat{d}(\mathbf{r}) = \sum_l w_l d_l. \tag{1}$$

A standard NeRF is supervised with the photometric loss between observed color and the expected color

$$L_{\text{photo}} = ||\hat{C}(\mathbf{r}) - C(\mathbf{r})||_2^2. \tag{2}$$

But in a setting with only $N < 20$ images to learn from, and frequent occurence of large untextured areas such as white walls, a NeRF is easily underconstrained

and does not learn the underlying geometry sufficiently. To improve geometric understanding, and to provide additional guidance during training, we incorporate depth supervision.

3.2 Depth Priors and Uncertainty

As monocular depth estimation is ambiguous, we require a measure of trust in the predictions. While many networks display promising depth prediction capabilities, generative models, especially denoising diffusion models, additionally let us reason about the depth generation process. DiffDP [11] is a state-of-the-art image-conditioned diffusion model designed for visual perception tasks like dense segmentation and depth estimation. For depth estimation, the network is trained to denoise noised ground truth depth maps using image features as conditioning for the denoising process. At inference, multi-resolution features are extracted from the input image and concatenated with random noise, where a lightweight decoder gradually denoises the input to generate a final depth prediction. Given an input image $\mathbf{I}$, DiffDP formulates its denoising process as

$$p_\theta(z_{0:T}|\mathbf{I}) = p_\theta(z_T) \prod_{t=1}^{T} p_\theta(z_{t-1}|z_t, \mathbf{I}), \tag{3}$$

where $z_t \sim \mathcal{N}(0, I)$ and z_0 corresponds to the models final depth estimate.

During this process, the network recursively updates its estimate. Pixels that the model is unsure about will be updated more often than others. We propose to use this measure as a proxy for uncertainty that is *not* focused on areas of high error, such as borders of objects, but instead captures uncertainty in the depth generation process itself.

To obtain the uncertainty of a depth estimate, $u(z_0)$, we compare the current estimate to the previous one at each time step t and report the count as:

$$c(z_{0:T}) = \frac{1}{T} \sum_{t=T,\dots,1} \mathbb{1}_{|z_t - z_{t-1}| \geq \tau}. \tag{4}$$

Let $\mathcal{M}(\cdot)$ denote a function that mirrors an image, then the uncertainty for an image x is

$$U(z_{0:T}|\mathbf{I}) = \frac{c(z_{0:T}|\mathbf{I}) + \mathcal{M}(c(z_{0:T}|\mathcal{M}(\mathbf{I})))}{2}. \tag{5}$$

We empirically found this measure to find extreme cases of uncertainty easily, while emphasizing somewhat uncertain areas less. To capture those areas as well, we augment the uncertainty map with a second dimension of uncertainty: We compare the final depth prediction of an image with the prediction of its mirrored image. The final uncertainty for an image I is then

$$u(z_{0:T}|\mathbf{I}) = U(z_{0:T}|\mathbf{I}) \cdot ||(z_0|\mathbf{I}) - \mathcal{M}(z_0|\mathcal{M}(\mathbf{I}))||_1. \tag{6}$$

An example uncertainty map is depicted in Figure 3, where DiffDP misinterprets the geometry of the electrical box, but highlights the same area as uncertain.

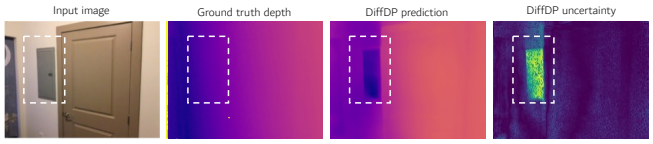


Fig. 3: Example of DiffDP depth prediction that misinterprets the depicted geometry. Although the predicted depth has errors, the uncertainty map is able to highlight areas of large errors. This allows us to tune down the depth loss in unreliable areas.

3.3 *EMD*-Guided Sampling

Our aim is to provide NeRF with depth *suggestions*. Even though we use a state-of-the-art depth prior, and an uncertainty measure that provides information about unreliable depth priors, we should still not assume that the prior is a perfect pseudo ground truth. Minimizing a norm between the NeRF-rendered depth and the depth prior, such as proposed in the concurrent work DäRF, would force the NeRF to approximate the depth priors irrespective of their accuracy. This could not only lead to a wrong understanding of the underlying scene, but also interfere with the RGB loss, which propagates gradient updates through shared layers. We thus seek to guide the ray termination distance distribution with the depth prior instead of directly supervising the NeRF-rendered depth.

A second motivation to leverage the distance distribution stems from NeRF's design. The volume density $\sigma(y)$ represents the differential probability that a ray terminates at y. We found empirically, that the learned ray termination distance distribution is rarely unimodal. So if we supervised the expected depth in Eq. 1 directly, we would collapse this probability into its expected value and lose this valuable information about the distribution. We thus supervise the ray termination distance distribution as proposed in SCADE [26] rather than supervising the predicted depth. To this end we normalize the predicted weights,w_i, to obtain a piecewise-constant probability density function along a ray: $\hat{w}_i = w_i / \sum_{j=1}^{l} w_j$.

We can then construct the cumulative density function (CDF) and sample new ray termination distances $\mathbf{y} = y_1, y_2, ...y_N$ through inverse transform sampling. The distribution of these samples can then be supervised in lieu of the less informative NeRF-sampled depth. But while SCADE models uncertainty in depth priors by providing 20 samples to choose from, the method still assumes that the chosen hypothesis is a perfect pseudo ground truth. We aim to relax this assumption. We instead propose to supervise the ray termination distance distribution with the Earth Mover's distance between the NeRF samples and the depth prior z_0:

$$L_{\text{EMD}} = \text{EMD}(\mathbf{y}, z_0). \tag{7}$$

We use Earth Mover's distance, as it naturally lends itself to compare probability densities or discrete histograms. It does not require satisfaction of assumptions such as that for two distributions P and Q, $Q(y) = 0 \implies P(y) = 0, \forall y$, which KL divergence would require. Earth Mover's distance also does not require

that $y_1, y_2, ...y_N$ are unimodally distributed, or that the error between NeRF-sampled ray termination distances and depth prior is normally distributed. Earth Mover's distance is therefore equipped to address complex uncertainties in our depth priors, such as those we empirically observe and that previous methods have not sufficiently captured. As EMD is not differentiable, we use Sinkhorn Divergence [8] to approximate the L_{EMD} during training.

3.4 Loss Weighting

Given the standard photometric loss and our novel depth guidance framework, we leverage the uncertainty we can capture for free from DiffDP's generation process to downweight pixels with especially uncertain depth prediction. Our work aims to provide a framework in which RGB-losses and depth losses complement each other: If we are certain about the depth of a pixel, we wish to upweight the depth loss. When we are unsure, we would rather rely on the photometric loss. Loosely inspired by Focal Loss [14], we define the total loss for a ray as

$$\mathcal{L} = (1+u)^{\gamma} L_{photo} + \lambda(1-u)^{\gamma} L_{\text{EMD}}, \tag{8}$$

where λ is a balancing weight and γ controls the impact of the uncertainties u. Opposed to Focal Loss which increases the weight of uncertain examples to force the model to learn hard cases, we apply less weight to uncertain pixels.

4 Results

In this section we evaluate the proposed method in detail, compare it to existing work, and explore the importance of our design choices.

4.1 Experimental Setup

We evaluate our method on three ScanNet [5] scenes as chosen by DDPrior, DäRF, and SCADE [22,25,26]. Each scene consists of 18-20 training images and 8 test images. To measure the generalization capability of our framework, we also evaluate on an additional dataset which we refer to as ScanNet+ containing additional scenes from ScanNet. Unless otherwise stated, all experiments are performed and averaged over the three standard ScanNet scenes. For evaluation, we use the standard photometric measures PSNR, SSIM, and LPIPS [35] as used in the original NeRF paper [15] and in all baselines we compare to. As we are especially interested in the ability to render accurate depths, we adapt the depth metrics used in DäRF, namely relative error (Rel) and RMSE [25].

We compare our work to several baselines that use depth-supervision for NeRF training in indoor environments. The most related baseline is DäRF [25], as it also leverage a frozen out-of-the-box pre-trained monocular depth prior to supervise NeRF training. SCADE is related to our work as well, but uses a

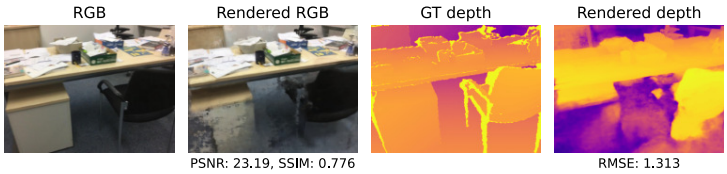


Fig. 4: Good RGB rendering quality does not imply good geometric understanding. In this test example, SCADE [26] accurately renders the image (PSNR and SSIM above average), while misinterpreting the geometry of the depicted scene (five-fold average RMSE). The model does not capture the cabinet below the desk in the depth map.

custom-trained depth prior, and does not evaluate the depth accuracy of their method. As SCADE provides pre-trained weights, we evaluate their method on all depth metrics. Additional baselines are a standard NeRF [15], DS-NeRF [7], and DDPrior [22], where we use the results reported by DDPrior. We report results on DDPrior using an out-of-domain prior as trained by SCADE, to replicate our setting where the depth network is trained on a different dataset.

4.2 Implementation Details

Our depth prior is the pretrained DiffDP network that was trained on the indoor depth dataset NYU [4]. During training we initialize the scale of the depth priors as 1, and learn the scale with a small learning rate. The uncertainty maps from DiffDP are derived once during construction of the prior and used at the beginning of NeRF training (normalized to $[0, 1]$). We use 1024 rays per batch, and sample 64 and 128 ray termination distances for the coarse and fine network, respectively. For the Earth Mover's distance we sample an additional 128 samples. We use the *geomloss* implementation of Sinkhorn with standard hyper-parameters. We employ dropout layers ($p = 0.1$) and weight decay ($\lambda_{wd} = 1e^{-6}$) for regularization. Further details of the prior construction, NeRF training, and evaluation can be found in the supplementary material.

4.3 Reconstruction Quality

In this section we evaluate our method's ability to render novel views, and, importantly, its ability to understand the geometric layout of a scene. We thus report both photometric and depth-based metrics. But first, we demonstrate that photometric and depth errors do not always go hand in hand. Observe in Figure 4, that a model can yield great photometric results, while misunderstanding the geometry. In the illustrated example, the cabinet is visible in two training images only and from similar angles. This leads to a misinterpretation of the scene such that the carpet appears to have a beige box pattern. Correcting the geometry in such cases does not improve rendering quality.

We compare our method on RGB-based and depth-based metrics in Table 1. Our method reduces all depth metrics of all baselines by at least 11%, and

Table 1: Experimental results on ScanNet [6]. D = Depth supervision during training: ✗ none, ✓ in-domain, ✢ out-of-domain. DiffDP is the depth prediction network we use as a prior. Our method dramatically improves NeRF's underlying scene geometry over initial depth priors while maintaing photometric reconstruction quality. Note that our method also outperforms DDPrior which has in-domain pretrained depth maps.

		RGB-based metrics			Depth-based metrics			
	D	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
DiffDP [11]	✢	-	-	-	0.100	0.032	0.261	0.128
NeRF [15]	✗	19.03	0.670	0.398	-	-	1.163	-
DS-NeRF [7]	✓	20.85	0.713	0.344	-	-	0.447	-
DDPrior [22]	✓	20.96	0.737	0.294	-	-	0.236	-
DDPrior* [22]	✢	19.29	0.695	0.368	-	-	0.474	-
SCADE** [26]	✢	21.54	0.732	**0.292**	0.086	0.030	0.252	0.118
DäRF [25]	✢	21.58	**0.765**	0.325	0.151	0.071	0.356	0.168
Ours	✢	**21.69**	0.737	0.373	**0.070**	**0.024**	**0.221**	**0.105**

* Trained by SCADE authors. ** Depth metrics evaluated by us based on SCADE's weights.

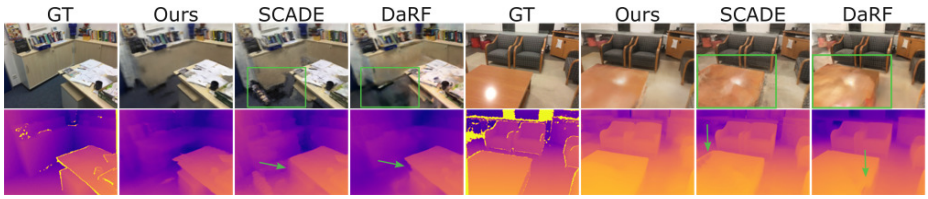


Fig. 5: Qualitative results of rendered RGB images and depth maps. Our method produces less artifacts in the left example and learns a better geometric representation of the table in the right example.

reduces the error of the most related work, DäRF, by 54%, 66%, 38%, and 38% on the four depth metrics. Our method also outperforms DDPrior [22] whose depth prior was trained in-domain. Our NeRF setup even improves the initial depth estimates of the depth prior DiffDP. Moreover, our method not only outperforms SCADE, but also avoids the need for a custom-trained depth prior. Training this prior requires evaluating the entire training set 20 times after each epoch to find the best hypothesis for each training example. Our training of SCADE's cIMLE-based prior takes more than one week on four 24GB NVIDIA Titan RTX GPUs.

When comparing RGB metrics to the closest baseline, SCADE, our method yields a 0.7% worse SSIM, while the PSNR is 0.7% better. Though our method yields an LPIPS that is on average worse than SCADE's reported result, when reproducing SCADE's results, some runs lead to an extremely high LPIPS above 0.5 (see Figure 9). Compared to DäRF, our SSIM is 3.7% worse, but all depth metrics are 37% to 66% improved. Overall, we highlight that when geometric

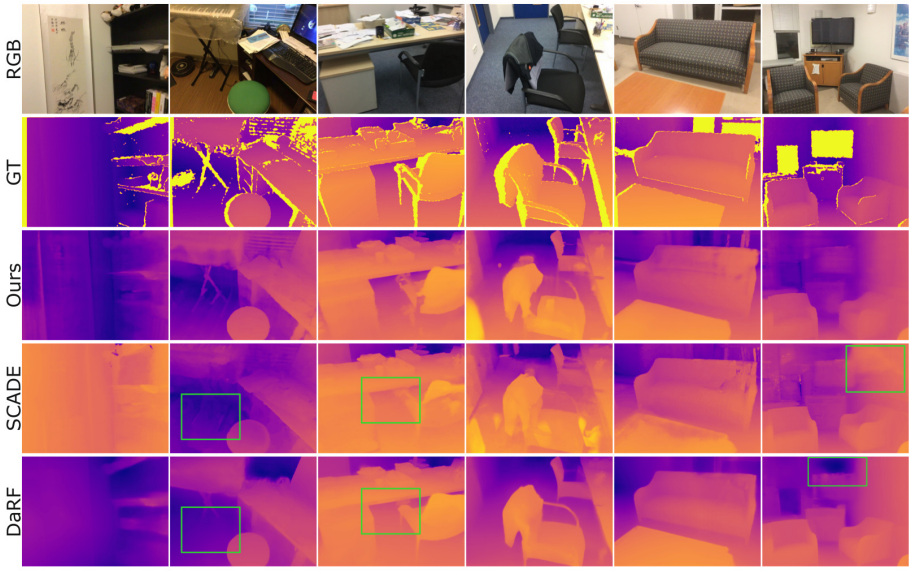


Fig. 6: Additional qualitative depth results. Our method produces less artifacts than SCADE, while more accurately capturing the underlying geometry than DäRF. The smoothness of DäRF's predictions is a result of it's strong supervision by the depth prior, but does not speak to the accuracy of the learned geometry.

consistency is desired, our method enables substantial gains in depth metrics while maintaining comparable RGB metrics.

We qualitatively compare results in Figure 5. In the illustrated examples, our method leads to less severe artifacts, and crisper and more accurate edges in RGB renderings and depth maps. Additional qualitative examples of depth maps are shown in Figure 6. Our model more accurately reconstruct the depicted layouts, like the area underneath the piano in the second image, or the area under the table in the third image. Interestingly, DäRF's depths are extremely crisp. However, the quantitative evaluation in Table 1 confirms that DäRF learns an inaccurate representation of the geometry. Its strong L_2-loss enforces the depth maps to look like their prior, although they do not correctly represent the scene.

To demonstrate the robustness of our method we report results on ScanNet+ that were not previously used by the baselines. We observe in Table 2, that our method reduces the RMSE of the baselines by at least 56%, while producing comparable photometric results. Please see the supplement for more results.

4.4 Role of Depth Supervision and Uncertainty

We wish to better understand how NeRFs leverage depth priors during training and conduct extensive ablations studies under various settings.

Table 2: To further demonstrate the generalization strength of our approach, we evaluate on a second dataset, ScanNet+ , which includes novel scenes not tested by prior work. We retrain all prior methods on these scenes and compare with ours.

	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓
SCADE [26]	20.57	0.687	0.369	0.396	1.032
DäRF [25]	22.01	**0.751**	**0.319**	0.745	1.765
Ours	**22.29**	0.722	0.391	**0.156**	**0.456**

Table 3: Ablation studies for the components of our method, varying depth priors and the objectives leveraging them. All reported results are on ScanNet scenes.

Reg.	z	EMD	u	PSNR ↑	SSIM ↑	LPIPS ↓	AbsRel ↓	RMSE ↓
✓				21.28	0.735	0.370	0.119	0.354
✓	✓			21.24	0.734	0.377	0.100	0.295
✓	✓	✓		**21.70**	0.736	**0.369**	0.071	0.222
✓	✓	✓	✓	21.69	**0.737**	0.377	**0.070**	**0.221**

(a) Impact of Regularization *Reg.*, the DiffDP prior z, the Earth Mover's distance loss EMD, and uncertainty u on the baseline NeRF model.

	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
DiffDP	**0.125**	**0.056**	**0.324**	**0.151**
DepthAnything	0.144	0.061	0.357	0.180

(b) Comparison of our depth prior DiffDP [11] and DepthAnything [30] evaluated on the training images used as input to our model.

With our loss (EMD)	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓
Ours w/ DiffDP‡	**21.70**	**0.736**	**0.369**	**0.222**
Ours w/ DepthAnything‡	21.62	0.732	0.378	0.290

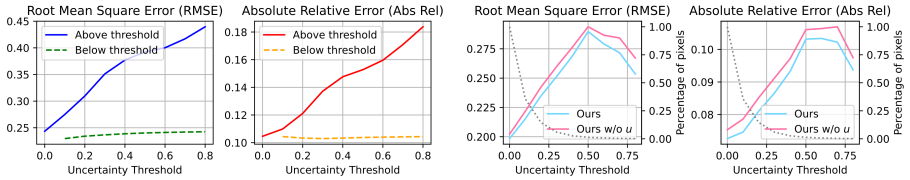
(c) Impact of the depth priors in (b) on our model. For both priors, our EMD objective allows the NeRF to learn smaller depth errors than the priors used for supervision yield.

With L_2-H loss	PSNR ↑	SSIM ↑	LPIPS ↓	RMSE ↓
Ours w/ DiffDP‡	**21.34**	**0.726**	**0.382**	**0.334**
Ours w/ DepthAnything‡	21.04	0.717	0.394	0.349

(d) Comparison of different depth priors under a L_2-objective applied to depth hypotheses (L_2-H loss ≡ space-carving loss [26] with only one depth hypothesis). ‡Without uncertainty u, as it cannot be used with DepthAnything prior.

In Table 3 (a) we evaluate the importance of design choices of our model compared to a basline NeRF model. We ablate Regularization *Reg.*, use of depth supervision z (by default with L_2 loss), use of EMD depth supervision instead of L_2 loss, and uncertainty weighting u. We observe that depth supervision plays a vital role in improving results, especially depth metrics. But depth alone (*Reg.* + z) does not yield satisfactory depth results. Replacing a naive L_2-loss with the EMD (*Reg.* + z + EMD + u) considerable reduces the depth error, highlighting the importance of depth guidance.

Although the impact of u is small when averaged across entire images, u makes a difference where it matters: in uncertain areas. In Figure 7 (a), we evaluate how well the uncertainty measure can distinguish reliable from unreliable pixels. We compute the depth error of DiffDP for pixels that have an uncertainty of $\geq t$ and compare it to more certain pixels with $u < t$ for thresholds $t = \{0, 0.1, ...0.8\}$. For $t = 0.5$, for instance, the depth error of uncertain pixels is 63% higher than the error of certain pixels. Our uncertainty measure thus reliably highlights pixels that should not be trusted. To evaluate its impact on our model, we ablate the use of u in Figure 7 (b) in a similar manner as the ablation in Table 3 (a); but instead of averaging over an entire image, we look at uncertain and certain regions of an image. We observe, that most pixels have a low uncertainty (64% of all pixels have an uncertainty of < 0.1), and for those pixels incorporating uncertainty only marginally (1.9%) helps. But for pixels with high uncertainty (>0.8), incorporating u improves the RMSE by 5.1%.



(a) (b)

Fig. 7: (a) Our uncertainty measure can identify areas in the DiffDP depth prior that we should not trust. We plot the error of the depth prior for pixels that are above or below a threshold for several thresholds. Both RMSE and Abs Rel are considerable higher in areas we predicted as uncertain. (b) We show the impact of incorporating our uncertainty u in our model. We plot depth errors for all pixels above a uncertainty threshold and ablate our method in comparison to a version without u. Especially for high uncertainty (> 0.6) incorporating u helps improve depth accuracy.

It is important to note that our model's ability to predict accurate depth does not solely stem from a good prior. We compare our prior to SCADE's LeReS-based prior [32] in Figure 8, where we adjusted the scale between prior and ground truth as ratio between their mean depths before calculating the RMSE. We make two interesting observations. First, sorting SCADE's 20 priors by accuracy, we find that the error increases rapidly. The additional priors introduce noise rather than adding beneficial information. This might explain why SCADE's training is at times unstable, and why learning accurate depth is so important for NeRFs. We can observe in Figure 9 that SCADE depends highly on the chosen random seed. The depth RMSE varies dramatically between runs. Comparing this to the PSNR, we observe that even when the depth prediction is widely off (depth error of 1.6 in scene 781), the NeRF can still produce decent RGB quality with a PSNR of at least 20. So while a NeRF can learn to produce realistic looking RGB images, it has no knowledge of the geometry (as seen in Figure 4). This not only prevents applications such as augmented reality in which a user would interact with the 3D world or when meshes need to be extracted for the scene, but also renders reported photometric results unreliable.

In Table 3 (b), we evaluate the initial accuracy of our DiffDP based prior and the prior from [30]. We show that DiffDP is a good choice for depth prior even in comparison to priors generated from large-scale pretraining like in DepthAnything [30]. Given this context, Table 3 (c) exemplifies our EMD-based objective's role in leveraging information from the priors. Irrespective of prior, our loss encourages a reduction in the initial prior's depth error (with our DiffDP prior, RMSE drops 31.4%, and with DepthAnything prior, RMSE drops 18.8%). Through EMD depth guidance, NeRFs can leverage these priors without enforcing them, leading to a smaller overall depth error. To further examine the impact of our objective, we construct a strong baseline under different depth priors that mimics the loss presented in [26] but with a single depth hypothesis.

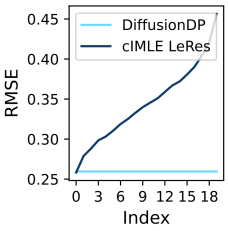


Fig. 8: Depth errors of cIMLE priors increase rapidly when arranged in ascending order.

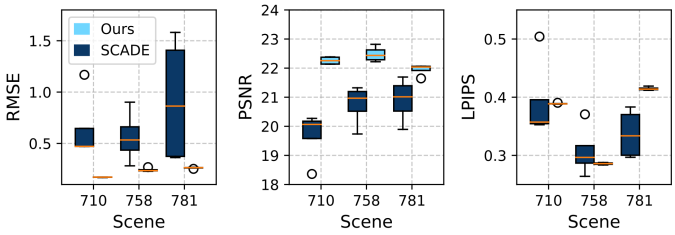


Fig. 9: We repeat the experiment in Table 1 with four random seeds and show the depth RMSE, PSNR, and LPIPS for each scene in ScanNet. While SCADE can sometimes produce accurate depths (low RMSE), the variability between runs is large. Our method produces consistently accurate depths.

Specifically, we use L_2 loss on depth hypotheses during NeRF training. We make two observations in Table 3 (d): First, depth metrics remain constant or even worsen in comparison to the initial depth priors from (b) under the L_2 objective. Secondly, our EMD-loss strongly outperforms the L_2 objective for both priors. These experiments highlight the importance of supervising NeRF with an EMD-based objective that can selectively incorporate information from the depth prior irrespective of the prior's strength.

5 Conclusions

While NeRFs can reliably render images from novel viewpoints, the underlying geometry is not always accurately learned. To improve the geometric understanding of NeRFs, we revisit depth supervision in NeRF training for the reconstruction of challenging indoor scenes. We present a simple, novel framework that leverages the depth estimates from pretrained diffusion models and their intrinsic notions of depth uncertainty. We show that noisy depth estimates coming from off-the-shelf depth estimation networks should not be used to directly supervise NeRF-rendered depths. Rather, weighted by uncertainty, the distribution of ray termination distances during NeRF optimization should be guided by depth priors through the Earth Mover's distance, allowing for selective supervision. Our method achieves strong empirical results and serves as an easy drop-in replacement for existing depth-supervised NeRFs.

Limitations and Future work: While different monocular depth priors can be leveraged in our EMD-guided framework, our uncertainty measure can only be obtained from diffusion-based depth networks. Future work should therefore include model-agnostic uncertainty measures. Additionally, the construction of the uncertainty map for the depth estimates can be sensitive to hyperparameters. An interesting direction for future work could be dynamically learning a threshold of uncertainty during the construction of the depth prior.

Acknowledgements

Thanks to the anonymous reviewers for their constructive feedback. This work was supported by the Isackson Family Foundation, the Stanford Head and Neck Surgery Research Fund, and the Stanford Graduate Fellowship.

References

1. Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., Schwager, M.: Vision-only robot navigation in a neural radiance world. IEEE Robotics and Automation Letters **7**(2), 4606–4613 (2022) 1
2. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023). https://doi.org/10.48550/ARXIV.2302.12288, https://arxiv.org/abs/2302.12288 2, 3
3. Blukis, V., Yoon, K.J., Lee, T., Tremblay, J., Wen, B., Kweon, I.S., Fox, D., Birchfield, S.: One-shot neural fields for 3d object understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW) on XRNeRF: Advances in NeRF for the Metaverse 2023. IEEE/CVF (2023) 1
4. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013) 9
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017) 3, 8
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017) 10
7. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022) 1, 2, 4, 9, 10
8. Feydy, J., Séjourné, T., Vialard, F.X., Amari, S.i., Trouvé, A., Peyré, G.: Interpolating between optimal transport and mmd using sinkhorn divergences. In: The 22nd International Conference on Artificial Intelligence and Statistics. pp. 2681–2690. PMLR (2019) 8
9. Guangcong, Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. IEEE/CVF International Conference on Computer Vision (ICCV) (2023) 4
10. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5865–5874. IEEE Computer Society, Los Alamitos, CA, USA (oct 2021). https://doi.org/10.1109/ICCV48922.2021.00583, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00583 4
11. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction (2023) 2, 3, 5, 6, 10, 12
12. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? (2017) 4
13. Li, Z., Chen, Z., Liu, X., Jiang, J.: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. Machine Intelligence Research **20**(6), 837–854 (2023) 3

14. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 3, 8
15. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 405–421. Springer International Publishing, Cham (2020) 1, 3, 5, 8, 9, 10
16. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S.M., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2022) 4
17. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 3
18. Prinzler, M., Hilliges, O., Thies, J.: Diner: Depth-aware image-based neural radiance fields. In: Computer Vision and Pattern Recognition (CVPR) (2023) 4
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 4
20. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021) 2
21. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020) 2
22. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Niebner, M.: Dense depth priors for neural radiance fields from sparse input views. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12882–12891. IEEE Computer Society, Los Alamitos, CA, USA (jun 2022). https://doi.org/10.1109/CVPR52688.2022.01255, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01255 1, 2, 3, 4, 8, 9, 10
23. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models (2023) 3
24. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 4
25. Song, J., Park, S., An, H., Cho, S., Kwak, M.S., Cho, S., Kim, S.: Därf: boosting radiance fields from sparse inputs with monocular depth adaptation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 68458–68470 (2023) 1, 2, 4, 8, 10, 12
26. Uy, M.A., Martin-Brualla, R., Guibas, L., Li, K.: Scade: Nerfs from space carving with ambiguity-aware depth estimates. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023) 1, 2, 4, 7, 8, 9, 10, 12, 13
27. Wang, Y., Long, Y., Fan, S.H., Dou, Q.: Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 431–441. Springer (2022) 1
28. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: 2021

IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5590–5599. IEEE Computer Society, Los Alamitos, CA, USA (oct 2021). https://doi.org/10.1109/ICCV48922.2021.00556, https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00556 4

29. Xie, Z., Geng, Z., Hu, J., Zhang, Z., Hu, H., Cao, Y.: Revealing the dark secrets of masked image modeling. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14475–14485. IEEE Computer Society, Los Alamitos, CA, USA (jun 2023). https://doi.org/10.1109/CVPR52729.2023.01391, https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.01391 3
30. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024) 3, 12, 13
31. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image (2023) 3
32. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 204–213 (2021). https://doi.org/10.1109/CVPR46437.2021.00027 13
33. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), http://arxiv.org/abs/2012.02190v3 4
34. Yuan, W., Gu, X., Dai, Z., Zhu, S., Tan, P.: Neural window fully-connected crfs for monocular depth estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3906–3915. IEEE Computer Society, Los Alamitos, CA, USA (jun 2022). https://doi.org/10.1109/CVPR52688.2022.00389, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.00389 3
35. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018) 8
36. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12786–12796 (2022) 1