

Digging into Depth Priors for Outdoor Neural Radiance Fields

Chen Wang

cw.chenwang@outlook.com

Tsinghua University, Baidu Research

Beijing, China

Jiadai Sun

Northwestern Polytechnical

University, Baidu Research

Beijing, China

Lina Liu

Zhejiang University, Baidu Research

Beijing, China

Chenming Wu*

RAL, Baidu Research

Beijing, China

Zhelun Shen*

RAL, Baidu Research

Beijing, China

Dayan Wu

Institute of Information Engineering,

Chinese Academy of Sciences

Beijing, China

Yuchao Dai

Northwestern Polytechnical

University

Xi'an, China

Liangjun Zhang

RAL, Baidu Research

Sunnyvale, CA, USA

ABSTRACT

Neural Radiance Fields (NeRFs) have demonstrated impressive performance in vision and graphics tasks, such as novel view synthesis and immersive reality. However, the shape-radiance ambiguity of radiance fields remains a challenge, especially in the sparse viewpoints setting. Recent work resorts to integrating depth priors into outdoor NeRF training to alleviate the issue. However, the criteria for selecting depth priors and the relative merits of different priors have not been thoroughly investigated. Moreover, the relative merits of selecting different approaches to use the depth priors is also an unexplored problem. In this paper, we provide a comprehensive study and evaluation of employing depth priors to outdoor neural radiance fields, covering common depth sensing technologies and most application ways. Specifically, we conduct extensive experiments with two representative NeRF methods equipped with four commonly-used depth priors and different depth usages on two widely used outdoor datasets. Our experimental results reveal several interesting findings that can potentially benefit practitioners and researchers in training their NeRF models with depth priors. Project page: <https://cwchenwang.github.io/outdoor-nerf-depth>

CCS CONCEPTS

- Computing methodologies → Computer vision; Computer graphics.

KEYWORDS

Neural Radiance Field, Depth Estimation, Depth Completion

*Corresponding authors ({wuchenming, shenzhelun}@baidu.com).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0108-5/23/10...\$15.00

<https://doi.org/10.1145/3581783.3612306>

ACM Reference Format:

Chen Wang, Jiadai Sun, Lina Liu, Chenming Wu, Zhelun Shen, Dayan Wu, Yuchao Dai, and Liangjun Zhang. 2023. Digging into Depth Priors for Outdoor Neural Radiance Fields. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3581783.3612306>

1 INTRODUCTION

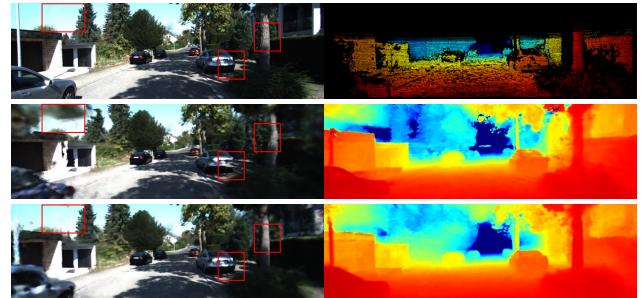


Figure 1: Image and depth map visualization of ground truth (top), trained with pure RGB (middle) and trained with monocular depth estimation (bottom), from a testing viewpoint. Even with a monocular depth (the quality is the worst compared to other depth priors), the view synthesis can be significantly improved for NeRF in terms of fewer floaters, and better preservation of object shapes (cars or trees) compared with using only RGB frames.

Novel view synthesis, *i.e.*, synthesizing photorealistic images at arbitrary viewpoints, is a long-standing task in computer vision and multimedia. Recently, neural radiance fields (NeRF) [35] and its variants have been a new selection for novel view synthesis and achieved impressive performance. Specifically, NeRF represents a 3D scene by a continuous function, which takes a pair of 3D position and 2D viewing direction as input to predict RGB color and volume density. This enables us to render an image using standard volume rendering equation [19]. The photorealistic renderings from NeRF

motivate a large amount of recent work in multimedia area [54, 55, 60, 63, 69, 73].

Despite the impressive development of NeRF, defining a reasonable underlying geometry from the radiance field is still an unresolved issue. To tackle this problem, several works use distance fields [30, 56] to explicitly define the geometry for NeRF frameworks. However, these methods require a large number of input images from different perspectives (normally 360-degree capturing is required) to successfully reconstruct an object. In outdoor scenes, there exist many foreground and background objects, and fully capturing all of them is a difficult task. Using geometry priors, in particular depth priors, to facilitate outdoor NeRF training is necessary and has been proven effective in previous work [34, 42, 58, 61]. Specifically, raw LiDAR depth, depth completion, and depth estimation are commonly used depth priors. Since they come at different costs and vary in accuracy, it is important to further investigate the criteria for selecting depth priors and the relative merits of them.

In this paper, we provide a comprehensive study and evaluation of fusing a different modal input, *i.e.*, depth priors, to outdoor neural radiance fields, covering common depth sensing technologies and most application ways in the all-time study of neural radiance fields. Specifically, depth sensing in outdoor scenes can be classified into two categories: (1) Active depth sensing: methods that employ optical devices to acquire depth. One major limitation is that the optical devices, *e.g.*, Light Detection and Ranging (LiDAR) are cumbersome and provide only sparse measurements. Depth completion is thus proposed to recover a dense depth map from a sparse one. (2) Passive depth sensing: methods directly infer depth from images, which is a much cheaper choice but sacrifices accuracy. Monocular and binocular depth estimation are two common methods. Experimentally, we select two representative NeRF methods and augment them with different depth supervision and different loss functions on two popular outdoor autonomous driving datasets: KITTI [13] and Argoverse [7]. As a result, we conclude the experimental results and have interesting findings as follows:

- **Density:** Even a very sparse depth supervision can significantly boost the view synthesis quality, and generally the denser the better.
- **Quality:** (1) Monocular depth is enough for sparse view inputs, which can even achieve comparable results with the ground truth depth supervision; (2) depth supervision is an option for dense view, *i.e.*, it is necessary if the corresponding application needs the employed NeRF to have a better geometry, such as 3D reconstruction.
- **Supervision:** Complex depth filtering and loss function is unnecessary in outdoor NeRF and directly cropping the sky area with MSE supervision is enough.

To the best of our knowledge, our work is the first quantitative and qualitative comparison of employing depth priors to outdoor neural radiance fields and we believe our findings would be helpful for practitioners and researchers to have a bigger picture of how to effectively incorporate depth priors in training outdoor NeRFs.

2 RELATED WORK

Neural Radiance Fields. Neural radiance fields (NeRFs) [36] demonstrate superior effectiveness in novel view synthesis by

predicting the per-point color and radiance of a 3D scene with a multi-layer perceptron (MLP). However, vanilla NeRF requires hours of optimization and assumes static scenes along with dense viewpoints. The following works have extended it in different aspects, *e.g.*, modeling dynamic and deformable scenes [40], super-resolution [55], sparse or imperfect poses [26], generalization to target scenes [68] and fast optimization [12]. To enable using NeRF in unbounded outdoor scenes, NeRF++ [74] introduces inverted sphere parametrization to handle unbounded scenes. MipNeRF-360 [2] re-parametrizes their scene coordinates with inverse-depth spacing, achieving evenly-spaced ray intervals in unbounded regions. In terms of large outdoor scenes, Block-NeRF [50] decomposes the scene into multiple blocks and trains a NeRF individually. Recent work [59, 67] validates the applicability of outdoor NeRF to autonomous driving simulation.

Depth Supervision for Neural Representations. Although vanilla NeRF only needs RGB images for training, when input viewpoints are sparse, the optimization can easily fall into local minima due to shape-radiance ambiguity. Additional depth supervision has been found to be useful in this scenario. DS-NeRF [10] firstly demonstrates the efficacy of depth information for NeRF with sparse inputs using the coarse point clouds from structure-from-motion. Dense depth priors [43] train an additional network for depth completion and uncertainty estimation, demonstrating the effectiveness of dense depth supervision. For street views, sparse LiDAR observations can be incorporated to supervise NeRF's geometry [42, 58, 62]. Apart from using ground truth depth, existing works also show that estimated monocular depth is also helpful in improving neural 3D reconstruction and view synthesis. MonoSDF [70] find monocular depth cues with off-the-shelf predictors can improve the quality and optimization time of neural surface reconstruction. NICER-SLAM [77] also integrates monocular depth to facilitate the mapping process in SLAM for indoor scenes. With the help of point clouds from monocular depth estimation, NoPe-NeRF [4] and Meuleman et al. [34] reconstruct NeRF and jointly estimate camera poses from a sequence of frames.

Depth Recovery. Current depth recovery technology can be roughly classified into three categories: (1) Depth completion. The goal of depth completion is to recover a dense depth map from a sparse one, *e.g.*, the depth acquired from LiDAR. Depth completion is divided into unguided methods and guided methods. Unguided methods [11, 53, 66] aim at directly completing a sparse depth map with a deep neural network. Guided methods [21, 28, 65] use RGB images to complete sparse depth maps to dense. Several strategies are proposed to improve the performance of depth completion, such as early fusion [32, 41], late fusion [29, 51], residual depth models [25, 27], and spatial propagation network based networks [9, 39]. (2) Monocular depth estimation. The goal of monocular depth estimation is to estimate a depth map from a single image. Early work mainly employs handcrafted feature [18, 44] to do monocular depth estimation, which often fails in complicated scenes. Currently, learning-based methods have shown their superiority, and encoder-decoder networks [5, 8, 23, 46] are the most commonly used architecture in this area. (3) Binocular depth estimation. The goal of binocular depth estimation, *i.e.*, stereo matching is to estimate a disparity/depth map from a pair of stereo images. It is a classic

task, and a well-known four-step pipeline [45] has been established. The early learning-based method [31, 71] mainly employs a convolutional network to replace one step of the traditional pipeline, *i.e.*, feature extraction. GCNet [20] is a breakthrough, which first proposes an end-to-end network to mimic the steps of the typical pipeline. Then, better feature extraction [24, 38], cost volume construction [16, 17], cost aggregation [6, 64, 72], and disparity computation [52, 76] are proposed by the follow-up methods to optimize the pipeline further.

3 DEPTH-SUPERVISED NeRF

As mentioned in Sec. 1, prior work has proved that depth priors are beneficial for NeRF training, especially in outdoor scenes, and multiple methods have been proposed to merge the depth prior into the NeRF framework. Tab. 1 classifies the existing depth-supervised NeRF methods, and two observations can be concluded from this table:

- Multiple depth priors have been applied in depth-supervised NeRF methods. However, all these methods only test one kind of depth prior and do not compare with other ones. Hence, the criteria for selecting depth priors and the relative merits of different priors have not been thoroughly investigated. Moreover, one available depth prior, binocular depth estimation, is missed in all existing work.
- Existing depth-supervised NeRF methods have proposed multiple loss functions to merge the depth prior into the NeRF framework. Similar to the former one, the relative merits of selecting different loss functions to use the depth priors is also an unexplored problem.

Hence, it is necessary to provide a comprehensive study and evaluation of employing depth priors to outdoor neural radiance fields. Specifically, we will give a classification of current depth-supervised NeRF methods and employed depth priors in this section.

3.1 Taxonomy of Depth-supervised NeRF

NeRF [35] represents a 3D scene as a continuous function that maps 3D positions $\mathbf{x} \in \mathbb{R}^3$ and 2D view directions $\mathbf{d} \in \mathbb{R}^2$ to radiance colors $\mathbf{c} \in \mathbb{R}^3$ and densities $\theta \in \mathbb{R}$. The function is typically parametrized with a MLP $f_\theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \theta)$. To render an image \mathbf{I} , we integrate the color along each camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ that shots from the camera center \mathbf{o} in direction \mathbf{d} with volume rendering:

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

where $T(t) = \exp(-\int_{t_n}^t \sigma(\mathbf{r}(t))dt)$ denotes the accumulated transmittance indicating the probability that a ray travels from t_n to t without hitting any particle. Similar to color, the depth map in NeRF can be rendered as follows:

$$\mathbf{D}_\theta(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))tdt. \quad (2)$$

Given a set of posed images $\mathcal{I} = \{I_i | i = 0, 1, \dots, N\}$, vanilla NeRF is optimized by comparing the mean-square error (MSE) between rendered images and their ground truth: $\mathcal{L}_{\text{MSE}}^{\text{rgb}} = \sum_i^N ||I_i - \hat{I}_i||_2^2$. Follow-up works augment NeRF training with additional depth information. To perform depth supervision, NeRF-based methods

Table 1: Classification of existing depth-supervised NeRF methods.

Methods	Depth priors	Loss Type
DS-NeRF [10]	SfM	KL
Urban-NeRF [42]	LiDAR	URF
S-NeRF [62]	Completion	L1
MonoSDF [70]	Mono	MSE
NICER-SLAM [77]	Mono	MSE
NoPe-NeRF [4]	Mono	L1
FEGR [58]	LiDAR	L1

Table 2: Taxonomy of depth priors. (·) denotes an optional selection.

Depth priors	LiDAR	Camera	Cost	Density
Raw LiDAR depth	✓	✗	high	sparse
Depth completion	✓	(✓)	high	dense
Monocular depth estimation	✗	✓	low	dense
Binocular depth estimation	✗	✓	middle	dense

first sample a batch of N_r rays, then compare the rendered depth and ground truth depth with different loss functions. Existing NeRF-based methods use depth supervision from two aspects: direct or indirect. Below we will introduce each category for more details.

Direct supervision. Direct supervisions directly compare the depth rendered by NeRF with that of the depth prior using supervision loss, including MSE and L1:

$$\mathcal{L}_{\text{MSE}}^d = \sum_i^{N_r} ||D(r_i) - \hat{D}(r_i)||^2, \quad (3)$$

$$\mathcal{L}_{\text{L1}} = \sum_i^{N_r} |D(r_i) - \hat{D}(r_i)|, \quad (4)$$

where $D(r_i)$ and $\hat{D}(r_i)$ are the predicted and ground truth depth of ray r_i . Both L1 and MSE loss are included in our experiment.

Indirect supervision. Indirect supervision uses depth prior to regularize the weights of NeRF, including the KL loss in DS-NeRF [10] and URF loss in Urban-NeRF [42]:

$$\mathcal{L}_{\text{KL}} = \sum_i^{N_r} \sum_k \log w_k \exp\left(-\frac{(t_k - D(r_i))^2}{2\hat{\sigma}^2}\right) \Delta t_k, \quad (5)$$

$$\mathcal{L}_{\text{URF}} = \sum_{t=t_n}^{D(r_i)-\epsilon} w(t)^2 + \sum_{t=D(r_i)-\epsilon}^{D(r_i)+\epsilon} (w(t) - \mathcal{K}_\epsilon(t - D(r_i)))^2, \quad (6)$$

where $D(r_i)$ and $\hat{D}(r_i)$ are the predicted and ground truth depth of ray r_i , w_k is the rendering weights of NeRF, t_k and Δt_k are the sampled points and distances of ray r_i , $w(t)$ is the weight corresponding to the point of distance t , $\epsilon(x)$ is a kernel that integrates to one (*i.e.*, a distribution) and has a bounded domain parameterized by ϵ . As the URF loss has no open-source implementation, we select the KL loss as the representative of indirect supervision.

3.2 Taxonomy of Depth Priors

As mentioned before, raw LiDAR depth, depth completion, monocular depth estimation, and binocular depth estimation are the main

depth priors employed in outdoor neural radiance fields. The taxonomy result of current depth priors is shown in Tab. 2 and we can conclude two observations from this table:

- Raw LiDAR depth and depth completion are LiDAR-based methods. One major limitation of these methods is that the employed LiDAR is cumbersome and provides only sparse measurements. On that basis, depth completion is proposed to recover a dense depth map from a sparse one.
- Monocular depth estimation and binocular depth estimation are camera-based methods. These methods can directly infer a dense depth from images, which is a much cheaper selection while making a large compromise in accuracy.

Below we will introduce each method in more detail.

Raw LiDAR depth. The raw LiDAR depth can be directly acquired from the employed optical devices, *i.e.*, LiDAR. As the LiDAR only provides sparse depth measurements, some methods also select to combine multi-frame of the point cloud [14, 33] to get a denser depth map.

Depth completion. Denote the input raw LiDAR depth as D_l and the corresponding image as I . The depth completion process can be represented as:

$$\begin{aligned} D_{\text{guide}} &= \delta(I, D_l), \\ D_{\text{unguide}} &= \delta(D_l), \end{aligned} \quad (7)$$

where D_{guide} and D_{unguide} denote unguided methods and guided methods, respectively. Currently, the latter can achieve higher accuracy with guidance from the image. Note that the image information is also the necessary input for NeRF. Hence, we select the guided method MFFNet [29] as our depth completion method.

Monocular depth estimation. Encoder-decoder networks [3, 22, 49] are the most commonly used architecture for this task. Let us define the input image as I . The whole monocular depth estimation process can be represented as:

$$D_{\text{mono}} = \varphi_d(\varphi_e(I)), \quad (8)$$

where φ_e denotes the encoder and φ_d denotes the decoder. We select a representative encoder-decoder network BTS [22] as our monocular depth estimation method. Note that this method is supervised and the outputs have the correct scale. For self-supervised or zero-shot pre-trained monocular depth estimation models, additional scale and drift should be estimated during training.

Binocular depth estimation. Feature extraction, cost volume construction, cost aggregation, and disparity computation are the typical pipeline of current deep stereo matching methods. Denote the input left and right images as I_l and I_r . The whole binocular depth estimation process can be represented as:

$$d = \eta(\delta(\partial(f_\theta(I_l), f_\theta(I_r)))), \quad (9)$$

where f_θ is the feature extraction network, ∂ is the cost volume construction network, δ is the cost aggregation network, and η is the disparity computation step. We select two representative stereo matching networks CFNet [47] and PCWNet [48] as our binocular depth estimation method.

4 EXPERIMENTS AND FINDINGS

In this section, we introduce the experiment settings and results of this paper. More details can be found in the [supplementary](#).

4.1 Dataset

KITTI [13] and Argoverse[7] are large datasets of real-world outdoor driving scenes. We evaluate and compare these methods on selected fragments from the KITTI odometry and Argoverse stereo sequences. In contrast to the object-centric datasets commonly used in NeRF, the vehicles in autonomous driving scenarios usually only move forward or turn. To reduce the influence of lighting changes and moving objects, we finally select five sequences from Seq 00, 02, 05, 06 in KITTI (125, 133, 175, 295, 320 frames) and three sequences from Argoverse (73, 72, 73 frames). Please refer to the supplemental material for details. For each sequence, we hold every 10 frames as the testing set, and the others are used for training. To verify the impact of sparse viewpoints, we simulate low-frequency imaging at 2.5 Hz. To this end, we select 25% of KITTI training data, *i.e.*, taking one for every 4. For the Argoverse dataset, we select 50% of training data, since its data logging frequency (5 Hz) is 1/2 of that in KITTI (10 Hz). We also make experiments on all the training data, *i.e.*, dense viewpoints. For the pose of the images, to avoid the inconsistency between the structure from motion (SfM) pose and the real depth scale, we use the poses provided by KITTI odometry and tracking poses from Argoverse.

4.2 Evaluation Metrics

4.2.1 Photorealistic Metrics. We use the common metrics in the novel view synthesis literature to compare the synthesized views at testing viewpoints with the ground truth: PSNR, SSIM [57] and LPIPS [75].

4.2.2 Depth Accuracy Metrics. Following previous work [15, 22], we use ABSREL (Mean Absolute Relative Error) and RMSE (Root Mean Squared Error) as our depth evaluation metrics.

4.3 Included NeRF Baselines

The original parametrization of NeRF can only deal with bounded or forward-facing scenes. Since we deal with unbounded real scenes, we select the following NeRF variants in our experiments.

NeRF++ [74] divides the unbounded scenes into two volumes: an inner unit sphere and an outer volume. Therefore, the volume rendering also consists of two parts. We render the depth in NeRF++ with an extended version of Eq. (2):

$$\begin{aligned} D(\mathbf{r}) &= \int_{t=0}^{t'} \sigma(\mathbf{r}(t)) t \cdot e^{-\int_{s=0}^s \sigma(\mathbf{r}(s)) ds} dt + \\ &\quad e^{-\int_{s=0}^{t'} \sigma(\mathbf{r}(s)) ds} \cdot \int_{t=t'}^{\infty} \sigma(\mathbf{r}(t)) t \cdot e^{-\int_{s=t'}^s \sigma(\mathbf{r}(s)) ds} dt, \end{aligned} \quad (10)$$

where $t \in (0, t')$ is inside the sphere and $t \in (t', \infty)$ is the unbounded area.

MipNeRF-360 [2] is an unbounded extension of MipNeRF [1], which proposes to use a contract function to parameterize the 3D Euclidean space within a ball of the radius of 2. For MipNeRF-360, we can directly render images and depth maps in the contracted space with Eq. (1) and Eq. (2).

Table 3: Quantitative comparison with selected methods on the KITTI dataset. The best and the second best results are shown in bold and underlined forms, respectively.

Methods	Depth Supervision	Dense					Sparse				
		PSNR↑	SSIM↑	LPIPS↓	RMSE↓	ABSREL↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	ABSREL↓
MipNeRF-360 [2]	RGB-Only	21.99	0.692	0.437	3.090	0.088	16.93	0.589	0.498	4.662	0.144
	GT Depth	<u>21.84</u>	<u>0.682</u>	<u>0.451</u>	<u>0.918</u>	0.032	19.14	0.630	0.474	<u>1.044</u>	0.040
	Depth Completion	21.51	0.670	0.467	0.818	0.026	<u>19.65</u>	<u>0.631</u>	0.482	1.030	0.032
	Stereo Depth	21.53	0.665	0.469	1.192	<u>0.030</u>	19.80	0.637	<u>0.475</u>	1.246	<u>0.034</u>
	Mono Depth	21.48	0.668	0.468	2.161	0.059	19.35	0.625	0.485	1.890	0.058
NeRF++ [74]	RGB-Only	20.29	0.520	0.585	48.638	3.917	17.60	0.535	0.562	56.253	4.960
	GT Depth	20.08	0.574	0.563	1.914	0.078	18.90	0.554	0.568	1.882	0.089
	Depth Completion	<u>20.15</u>	0.576	0.560	2.618	0.102	18.90	<u>0.553</u>	<u>0.569</u>	<u>2.022</u>	<u>0.094</u>
	Stereo Depth	20.10	<u>0.575</u>	0.560	<u>1.934</u>	<u>0.087</u>	18.85	0.550	0.574	2.061	0.100
	Mono Depth	19.87	0.566	0.567	2.256	0.092	18.74	0.548	0.574	2.670	0.110
Instant-NGP [37]	RGB-Only	20.51	0.630	<u>0.460</u>	9.575	0.507	15.44	0.499	0.536	15.011	0.793
	GT Depth	21.31	0.650	0.444	1.571	<u>0.052</u>	18.53	0.586	0.469	1.751	<u>0.060</u>
	Depth Completion	20.90	<u>0.632</u>	0.470	<u>1.661</u>	0.050	18.62	<u>0.576</u>	<u>0.492</u>	<u>1.833</u>	0.059
	Stereo Depth	<u>20.93</u>	0.629	0.472	1.830	0.057	<u>18.60</u>	0.574	0.493	1.984	0.064
	Mono Depth	20.59	0.617	0.483	2.679	0.085	18.17	0.557	0.502	2.868	0.096

Table 4: Quantitative evaluations of depth prior on the KITTI dataset with the selected sequences. Please refer to the previous work [15, 22] for the specific definition of evaluation metrics.

Methods	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	ABSREL↓	Sq Rel↓	RMSE↓	RMSE log↓	Density
Monocular Estimation	0.970	0.997	0.999	0.058	0.156	2.020	0.085	100%
Stereo Matching	0.996	0.998	0.999	0.016	0.035	1.080	0.040	100%
Stereo Matching_confidence	0.999	0.999	0.999	0.014	0.016	0.71	0.025	92.28%
Depth Completion	0.998	1.000	1.000	0.010	0.015	0.622	0.020	100%

Instant-NGP [37] proposes a novel scene representation that bounds an actual scene into an axis-aligned bounding box and uses a small neural network augmented by a multi-resolution hash table of trainable feature vectors whose values are optimized through stochastic gradient descent. The features are further mapped to color and density. Since it still uses standard volume rendering, the depth in Instant-NGP can be similarly rendered as in vanilla NeRF (Eq. (2)).

4.4 Evaluation Results and Comparisons

In this section, we conduct experiments on both Argoverse and KITTI datasets to verify the relative merits of employing different depth priors. Below we describe each dataset's result in more detail.

KITTI The qualitative depth-supervised NeRF results and corresponding depth prior quality evaluation can be found in Tab. 3 and 4. As shown in Tab. 4, the performance gap between different depth priors is large. Specifically, depth completion has the highest accuracy then goes with binocular depth estimation and monocular depth estimation. Below we will further analyze the relative merits of employing different depth priors to dense and sparse views.

(1) **Sparse view.** We first discuss the experiment result on the sparse view setting. As shown, NeRF trained with pure RGB suffers from heavy shape-radiance ambiguity (Fig. 2), so the view synthesis quality at novel viewpoints degrades significantly. In this case, the importance of depth information is evident, and we can see from

Tab. 3 that any type of depth can be conducive and greatly improve the synthesized views. Taking MipNeRF-360 as an example, we can see 11.55%~14.49% photorealistic metrics improvement (PSNR) and 59.72%~77.78% depth accuracy metrics improvement (ABSREL) with any type of depth prior. Moreover, we can observe that the photorealistic performance gain of using different depth prior is close even if the depth quality gap between different depth priors is large. That is, we can employ the cheapest depth prior (*i.e.*, monocular depth estimation) to achieve similar performance improvement with the costly depth prior (*e.g.*, the ground truth depth collected by LiDAR). A similar situation can be observed on Instant-NGP. Hence, we can get our **finding 1: Monocular depth is enough for sparse view.** Our first counter-intuitive finding is that using monocular depth estimation can significantly improve the quality of NeRF and even achieve comparable results with the ground truth depth supervision in sparse view. Generally, the monocular depth estimation is a cheaper selection and does not need additional equipment, *e.g.*, LiDAR. Thus, monocular depth estimation is a better selection in sparse view and binocular depth estimation is also an option if you need better depth map quality.

(2) **Dense view.** We then discuss the experiment result on the dense view setting. As shown, depth supervision is also helpful for the depth accuracy metrics. Taking MipNeRF-360 as an example, we can see 32.96%~70.45% depth accuracy metrics improvement (ABSREL) with any type of depth prior. That is the depth prior

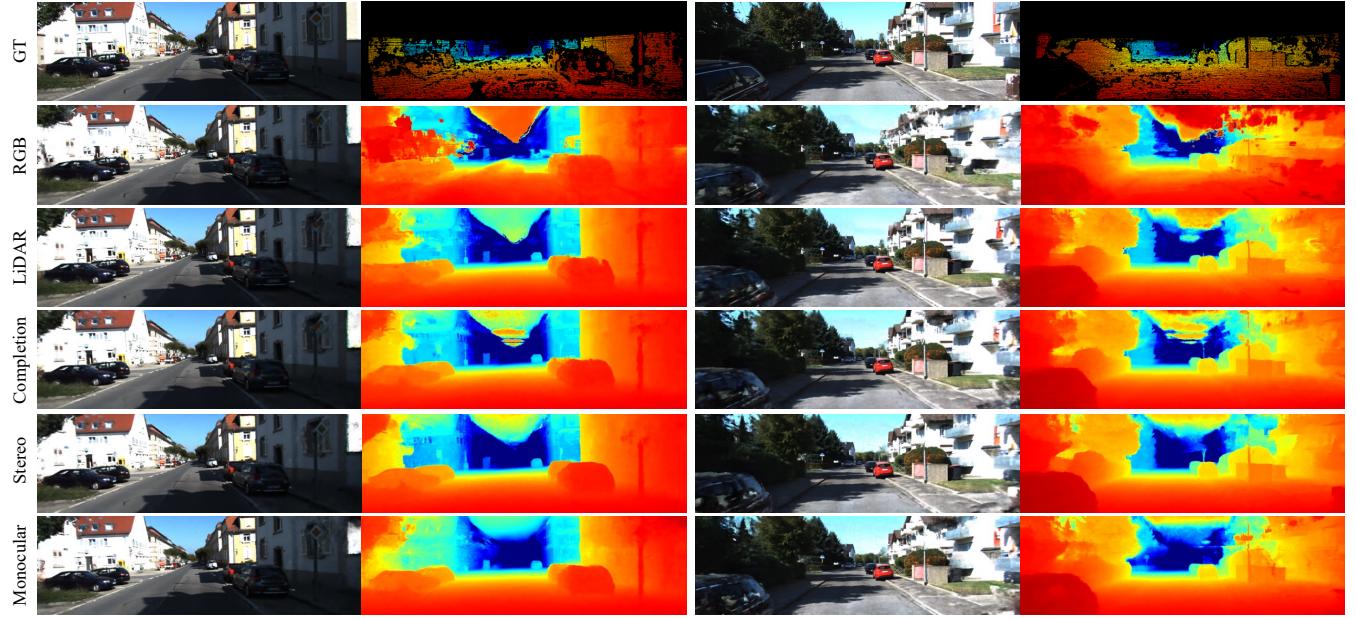


Figure 2: Qualitative results on the KITTI dataset with MipNeRF-360 with sparse viewpoints. Compared with training with RGB, adding depth supervision improves quality significantly. Better viewed zoomed-in and in-color.

Table 5: Evaluation and comparison of MipNeRF-360 and Instant-NGP on the Argoverse dataset. The best and the second best results are shown in bold and underlined forms, respectively.

Methods	Depth Supervision	Dense						Sparse					
		PSNR↑	SSIM↑	LPIPS↓	RMSE↓	ABSREL↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	ABSREL↓		
MipNeRF-360 [2]	RGB-Only	29.35	0.855	0.446	6.113	0.120	25.81	0.829	0.468	6.971	0.139		
	GT Depth	28.78	0.846	0.458	2.251	0.044	<u>28.01</u>	0.840	0.462	2.443	0.048		
	Stereo Depth	28.32	0.837	0.470	<u>4.271</u>	<u>0.064</u>	27.72	0.833	0.471	<u>4.310</u>	<u>0.067</u>		
	Mono Depth	28.58	0.841	0.466	4.601	0.093	28.04	0.836	0.468	4.868	0.093		
Instant-NGP [37]	RGB-Only	28.07	0.847	0.450	13.478	0.493	22.18	0.816	0.494	17.439	0.593		
	GT Depth	28.92	0.847	0.450	<u>1.804</u>	0.045	27.38	0.834	0.460	1.881	0.048		
	Stereo Depth	<u>28.32</u>	0.839	0.460	<u>5.613</u>	<u>0.090</u>	<u>27.10</u>	0.828	<u>0.467</u>	<u>5.843</u>	<u>0.097</u>		
	Mono Depth	28.31	0.838	0.466	6.083	0.122	26.97	<u>0.829</u>	0.471	6.643	0.132		

is still essential for the radiance field to obtain a reasonable underlying geometry. On the other hand, the performance gain in photorealistic metrics is not so noteworthy (Instant-NGP) or even causes a performance drop in some methods (MipNeRF-360). We attribute the drop in performance to inconsistent optimization directions for depth and RGB under contraction functions since one of the main differences between Instant-NGP and MipNeRF-360 is the usage of unbounded contraction. As a result, we can get our **finding 2: depth supervision is an option for dense view**. Our second interesting finding is that using depth supervision can achieve significant geometry improvement and trivial photorealistic metrics improvement in dense view. Thus, depth supervision is an option in dense view. It is still necessary if the corresponding application needs the employed NeRF to have a better geometry, such as reconstruction and potentially relighting, shadowing, etc.

Argoverse We also experimented on the Argoverse dataset to further verify our claim. Please note that the Argoverse dataset does not provide the depth completion task, so we do not include this setting in the experiment. The qualitative results can be found in Tab. 5. As shown, the view synthesis quality at novel viewpoints also degrades significantly in sparse views, and any depth can greatly improve the synthesized views. We also observe significant depth accuracy metrics improvement and trivial photorealistic metrics for the dense view setting. The situation is the same with the KITTI dataset, which further supports the validity of our findings 1&2.

4.5 Ablation Study

Although experimental results in Sec. 4.5 has demonstrated the relative merits of employing different depth priors, there are still many underlying settings that can reveal more profound findings, such as the impact of different depth densities, depth range, confidence level, depth loss function, etc. In this section, we conduct

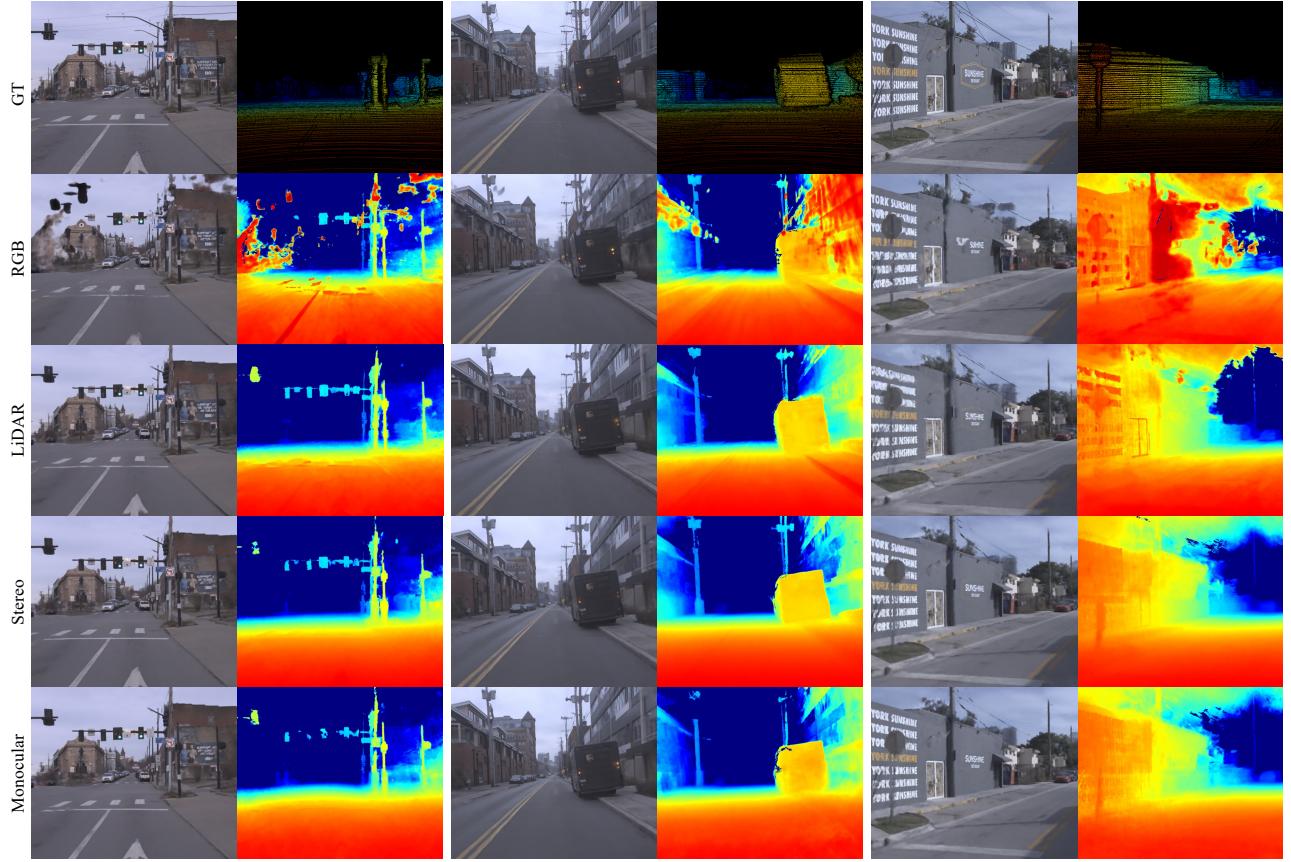


Figure 3: Qualitative results on the Argoverse dataset with MipNeRF-360 with sparse viewpoints (the GT depth is dilated with a 3×3 kernel for better visualization, which is extremely sparse). Compared with training with RGB, adding depth supervision improves quality significantly. Better viewed zoomed-in and in-color.

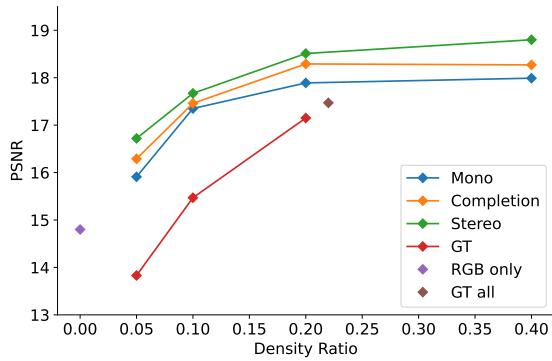


Figure 4: The relationship between PSNR and the density ratio of depth supervision under different types of depth priors.

several ablation experiments on one sequence of the KITTI dataset with MipNeRF-360 and sparse view setting to further explore those factors.

4.5.1 Density. Density is the main factor that differentiates the LiDAR ground truth from other depth priors, e.g. the density of LiDAR ground truth on the KITTI dataset is nearly 20%, and the other depth priors is 100%. We conduct two ablation experiments to further investigate the influence of different densities.

GT Masking It is interesting that monocular depth estimation can achieve on-par or even better performance than ground truth LiDAR in photorealistic metrics. This is likely because monocular depth is much denser than LiDAR, which only provides valid depth gt for around 20% of pixels. To verify this hypothesis, we try to use the invalid location of ground truth LiDAR to mask out the monocular depth estimation result and make sure the two depth priors have the same density. The results show that the photorealistic performance of using monocular depth drops to be close to or inferior to using ground truth depth after masking. This indicates that even imprecise dense supervision can still be more beneficial than sparse one in sparse views for novel view synthesis.

Depth Density To further evaluate the influence of supervision sparsity, we also test different depth densities by iteratively removing a fixed proportion of pixels from the original depth priors. The results in Fig. 4 show that a tiny amount (5%) of depth supervision

is enough to improve the performance of NeRF, and a denser depth supervision can achieve a more considerable gain. Specifically, the improvement is more significant before 20%, and the results become stable after enough supervision (40%). Moreover, we also observe that even with randomly selected less than 20% of monocular depth supervision (or depth completion and stereo depth), the results already surpass using all the ground truth (22%), which indicates that imprecise while covering a wider range of areas supervision (e.g., monocular depth estimation) is better than accurate while limited in central regions supervision (e.g., LiDAR) in sparse view. The GT Masking experiment further supports our claim, i.e., a sparse and covering the same region monocular depth estimation cannot beat the ground truth LiDAR supervision.

Considering the above experiments, we can obtain our **finding 3: The denser the better**. Our third interesting finding is that even a very sparse depth supervision can significantly boost the quality of novel view synthesis in sparse view, and generally, we observe that the denser the depth, the better quality of novel view synthesis we can obtain.

4.5.2 Depth Loss Type. Apart from MSE, we also test L1 and KL losses. As shown in Tab. 6, there are no significant performance differences between L1 and MSE loss. Moreover, KL loss is significantly worse than MSE, possibly because it is an indirect loss function and has a strong constraint in NeRF optimization, which would cause adverse effects if the depth estimation is not accurate enough.

4.5.3 Depth Filtering. The quality of depth varies widely among different methods. Generally, the quality degrades from completion to stereo, and finally to monocular (refer to Tab. 4). To address the unstable presentation of depth quality in many downstream tasks, there exist a few approaches to filter certain depth values and only use a subset of them. In this study, we choose two simple and widely-used depth filtering approaches, i.e., threshold clipping and confidence-based filtering, to investigate the influence.

Threshold Clipping In Tab. 5 and Tab. 3, we use depth prediction from different methods directly (cropping the sky area). Generally, the background area (far location) is less accurate than the foreground area (close location). Here, we test the impact of filtering out predictions farther than the threshold, i.e., large than a threshold s . In our experiments, we set s as 40m and 80m. The results are shown in Tab. 6. We can see that there are no significant differences in both image and depth metrics before and after clipping with $s = 80m$. For $s = 40m$, the image and depth metrics drop a little, likely because it lost the depth information of the background area, although it is less accurate than the foreground area.

Confidence-based Filtering Confidence estimation is also a common operation in both depth estimation and depth-supervised NeRF. Here, we test such confidence filters in binocular depth estimation as an example. Specifically, the confidence filters are implemented by the uncertainty estimation proposed in CFNet. The corresponding depth quality evaluation result is shown in Tab. 4. From Tab. 6, we can see that similar to threshold clipping, after filtering, the performance has no notable change.

Table 6: Ablation study with GT masking, threshold clipping, confidence filtering, and loss function.

Experiments	Factor	Depth Type	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	ABSREL↓
-	RGB-Only	-	14.80	0.475	0.551	4.569	0.153
GT Masking	-	LiDAR	17.47	0.542	0.507	1.173	0.045
	Yes	Mono	17.11	0.535	0.512	2.345	0.076
	No	Mono	17.97	0.542	0.510	2.383	0.073
Threshold Clipping	-	Mono	17.97	0.542	0.510	2.383	0.073
	40m	Mono	17.60	0.541	0.509	2.470	0.071
	80m	Mono	18.18	0.542	0.510	2.390	0.073
Confidence Filtering	No	Stereo	18.87	0.562	0.501	1.349	0.0405
	Yes	Stereo	18.85	0.565	0.495	1.467	0.0424
Loss Function	MSE	Mono	17.97	0.542	0.510	2.383	0.073
	L1	Mono	17.91	0.519	0.550	2.543	0.075
	KL	Mono	16.55	0.526	0.515	2.487	0.076

Finding 4: Simple loss function and depth filtering are enough. The above three ablation studies lead to our fourth finding: Complex depth filtering and loss function is unnecessary in outdoor NeRF and directly cropping the sky area (point at infinity) with MSE supervision is enough.

4.6 Discussion

Similar to DS-NeRF [10], we also conduct experiments by training Instant-NGP for 30 epochs and found that additional depth **accelerates convergence speed** and is also **beneficial in the extreme few-shot setting**. In the sparse setting, training with depth surpasses the training with only RGB with 30 epochs, even at the first epoch. For extremely sparse views, we further select 1/8, 1/16 of the input views for training in the Kitti dataset. The resulting PSNR/LPIPS is 13.1/0.57 (RGB Only), 16.4/0.52 (Monocular depth) for 1/8, and 11.6/0.61 (RGB Only), 13.0/0.57 (Monocular depth) for 1/16. Using other depth supervision is even better than using monocular depth.

5 CONCLUSION

This paper presents the first in-depth study and evaluation of employing depth priors to outdoor neural radiance fields, covering all common depth sensing technologies and most application ways. As a result, we conclude the experimental results and have interesting findings as follows: (1) **Density**: Even a very sparse depth supervision can significantly boost the view synthesis quality, and generally, the denser, the better; (2) **Quality**: (a) Monocular depth is enough for the sparse view, which can even achieve comparable results with the ground truth depth supervision. (b) depth supervision is an option for dense view, i.e., the depth supervision is necessary if the corresponding application needs the employed NeRF to have a better geometry; (3) **Supervision**: Complex depth filtering and loss function is unnecessary in outdoor NeRF and directly cropping the sky area with MSE supervision is enough. We believe these findings can potentially benefit practitioners and researchers in training their NeRF models with depth priors.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62106258 and 62271410.

REFERENCES

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proc. of the IEEE/CVF Int'l. Conf. on Computer Vision (ICCV)*.
- [2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. 2021. Adabins: Depth estimation using adaptive bins. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 4009–4018.
- [4] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. 2022. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. *arXiv preprint arXiv:2212.07388* (2022).
- [5] Yuanzhouhan Cao, Yidong Li, Haokui Zhang, Chao Ren, and Yifan Liu. 2021. Learning Structure Affinity for Video Depth Estimation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 190–198.
- [6] Jia-Ren Chang and Yong-Sheng Chen. 2018. Pyramid stereo matching network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5410–5418.
- [7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. 2019. Argoverse: 3d tracking and forecasting with rich maps. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 8748–8757.
- [8] Zhi Chen, Xiaoqing Ye, Liang Du, Wei Yang, Liusheng Huang, Xiao Tan, Zhenbo Shi, Fumin Shen, and Errui Ding. 2021. AggNet for Self-supervised Monocular Depth Estimation: Go An Aggressive Step Furthe. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1526–1534.
- [9] Xinjing Cheng, Peng Wang, and Ruigang Yang. 2018. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*. 103–119.
- [10] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [11] Abdela Rahmen Eldesokey, Michael Felsberg, Karl Holmquist, and Michael Persson. 2020. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 12014–12023.
- [12] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance Fields without Neural Networks. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [13] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3354–3361.
- [15] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3828–3838.
- [16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2495–2504.
- [17] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. 2019. Group-wise correlation stereo network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3273–3282.
- [18] Jose L Herrera, Carlos R Del-Blanco, and Narciso Garcia. 2018. Automatic depth extraction from 2D images using a cluster-based learning framework. *IEEE Trans. on Image Processing (TIP)* 27, 7 (2018), 3288–3299.
- [19] James T Kajiya and Brian P Von Herzen. 1984. Ray tracing volume densities. *ACM SIGGRAPH computer graphics* 18, 3 (1984), 165–174.
- [20] Alex Kendall, Hayk Martirosyan, Saumitra Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. 2017. End-to-end learning of geometry and context for deep stereo regression. In *IEEE International Conference on Computer Vision (ICCV)*. 66–75.
- [21] Md Fahim Faysal Khan, Nelson Daniel Troncoso Aldas, Abhishek Kumar, Siddharth Advani, and Vijaykrishnan Narayanan. 2021. Sparse to dense depth completion using a generative adversarial network with intelligent sampling strategies. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5528–5536.
- [22] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326* (2019).
- [23] Rui Li, Xiantuo He, Yu Zhu, Xianjun Li, Jinqiu Sun, and Yanning Zhang. 2020. Enhancing self-supervised monocular depth estimation via incorporating robust constraints. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3108–3117.
- [24] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. 2019. Stereo matching using multi-level cost volume and multi-scale feature constancy. In *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*.
- [25] Yiyi Liao, Lichao Huang, Yue Wang, Sarah Kodagoda, Yinan Yu, and Yong Liu. 2017. Parse geometry from a line: Monocular depth estimation with partial laser observation. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*. 5059–5066.
- [26] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5741–5751.
- [27] Lina Liu, Yiyi Liao, Yue Wang, Andreas Geiger, and Yong Liu. 2021. Learning steering kernels for guided depth completion. *IEEE Trans. on Image Processing (TIP)* 30 (2021), 2850–2861.
- [28] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. 2021. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 2136–2144.
- [29] Lina Liu, Xibin Song, Jiadai Sun, Xiaoyang Lyu, Lin Li, Yong Liu, and Liangjun Zhang. 2023. MFF-Net: Towards Efficient Monocular Depth Completion with Multi-modal Feature Fusion. *IEEE Robot. Automat. Lett. (RA-L)* (2023).
- [30] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. 2023. NeuralUDF: Learning Unsigned Distance Fields for Multi-view Reconstruction of Surfaces with Arbitrary Topologies. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [31] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. 2016. Efficient deep learning for stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5695–5703.
- [32] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. 2019. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*. 3288–3295.
- [33] Moritz Menze and Andreas Geiger. 2015. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3061–3070.
- [34] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. 2023. Progressively Optimized Local Radiance Fields for Robust View Synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*.
- [36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* (2021).
- [37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. on Graphics (TOG)* (2022).
- [38] Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. 2019. Multi-level context ultra-aggregation for stereo matching. In *IEEE conference on computer vision and pattern recognition (CVPR)*. 3283–3291.
- [39] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. 2020. Non-local spatial propagation network for depth completion. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*. 120–136.
- [40] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.
- [41] Chao Qu, Ty Nguyen, and Camillo Taylor. 2020. Depth completion via deep basis fitting. In *Proc. of the IEEE Winter Conf. on Applications of Computer Vision (WACV)*. 71–80.
- [42] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. 2022. Urban radiance fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
- [43] Barbara Roessel, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12892–12901.
- [44] Ashutosh Saxena, Min Sun, and Andrew Y Ng. 2008. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)* 31, 5 (2008), 824–840.
- [45] Daniel Scharstein and Richard Szeliski. 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision (IJCV)* 47, 1–3 (2002), 7–42.
- [46] Guibao Shen, Yingkui Zhang, Jialu Li, Mingqiang Wei, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. 2021. Learning regularizer for monocular depth estimation with adversarial guidance. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5222–5230.
- [47] Zhelun Shen, Yuchao Dai, and Zhibo Rao. 2021. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition.* 13906–13915.
- [48] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. 2022. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 280–297.
 - [49] Minsoo Song, Seokjae Lim, and Wonjun Kim. 2021. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE transactions on circuits and systems for video technology* 31, 11 (2021), 4381–4393.
 - [50] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. 2022. Blocknerf: Scalable large scene neural view synthesis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
 - [51] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. 2020. Learning guided convolutional network for depth completion. *IEEE Trans. on Image Processing (TIP)* 30 (2020), 1116–1129.
 - [52] Stepan Tulyakov, Anton Ivanov, and Francois Fleuret. 2018. Practical deep stereo (pds): Toward applications-friendly deep stereo matching. *Advances in neural information processing systems* 31 (2018).
 - [53] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. 2017. Sparsity invariant cnns. In *Proc. of the Intl. Conf. on 3D Vision (3DV)*, 11–20.
 - [54] Chen Wang, Angtian Wang, Junbo Li, Alan Yuille, and Cihang Xie. 2023. Benchmarking robustness in neural radiance fields. *arXiv preprint arXiv:2301.04075* (2023).
 - [55] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. 2022. NeRF-SR: High Quality Neural Radiance Fields using Super-sampling. In *Proc. of the ACM Intl. Conf. on Multimedia. (MM)*.
 - [56] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*.
 - [57] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
 - [58] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. 2023. Neural Fields meet Explicit Geometric Representation for Inverse Rendering of Urban Scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.
 - [59] Chennming Wu, Jiadai Sun, Zhelun Shen, and Liangjun Zhang. 2023. MapNeRF: Incorporating Map Priors into Neural Radiance Fields for Driving View Simulation. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*.
 - [60] Zijin Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. 2022. DoF-NeRF: Depth-of-Field Meets Neural Radiance Fields. In *Proc. of the ACM Intl. Conf. on Multimedia. (MM)*, 1718–1729.
 - [61] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. 2023. S-NeRF: Neural Radiance Fields for Street Views. In *Proc. of the Int. Conf. on Learning Representations (ICLR)*.
 - [62] Ziyang Xie, Junge Zhang, Wenye Li, Feihu Zhang, and Li Zhang. 2023. S-NeRF: Neural Radiance Fields for Street Views. *arXiv preprint arXiv:2303.00749* (2023).
 - [63] Wenpeng Xing and Jie Chen. 2022. MVSPlenOctree: Fast and Generic Reconstruction of Radiance Fields in PlenOctree from Multi-view Stereo. In *Proc. of the ACM Intl. Conf. on Multimedia. (MM)*, 5114–5122.
 - [64] Haofei Xu and Juyong Zhang. 2020. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1959–1968.
 - [65] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. 2019. Depth completion from sparse lidar data with depth-normal constraints. In *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2811–2820.
 - [66] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. 2019. Hierarchical deep stereo matching on high-resolution images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 5515–5524.
 - [67] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. 2023. UniSim: A Neural Closed-Loop Sensor Simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1389–1399.
 - [68] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.
 - [69] Xianggang Yu, Jiapeng Tang, Yipeng Qin, Chenghong Li, Xiaoguang Han, Linchao Bao, and Shuguang Cui. 2022. PVSeRF: joint pixel-, voxel- and surface-aligned radiance field for single-image novel view synthesis. In *Proc. of the ACM Intl. Conf. on Multimedia. (MM)*, 1572–1583.
 - [70] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665* (2022).
 - [71] Jure Zbontar, Yann LeCun, et al. 2016. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.* 17, 1 (2016), 2287–2318.
 - [72] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. 2019. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 185–194.
 - [73] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. 2022. VMRF: View Matching Neural Radiance Fields. In *Proc. of the ACM Intl. Conf. on Multimedia. (MM)*, 6579–6587.
 - [74] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).
 - [75] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 586–595.
 - [76] Youmin Zhang, Yimin Chen, Xiao Bai, Jun Zhou, Kun Yu, Zhiwei Li, and Kuiyuan Yang. 2019. Adaptive Unimodal Cost Volume Filtering for Deep Stereo Matching. In *arXiv preprint*.
 - [77] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. 2023. NICER-SLAM: Neural Implicit Scene Encoding for RGB SLAM. *arXiv preprint arXiv:2302.03594* (2023).