GeoDepth: From Point-to-Depth to Plane-to-Depth Modeling for Self-Supervised Monocular Depth Estimation

Haifeng Wu, Shuhang Gu, Lixin Duan, Wen Li*
University of Electronic Science and Technology of China
{haifengwu205, shuhanggu, lxduan, liwenbnu}@gmail.com

Abstract

Self-supervised monocular depth estimation has long been treated as a point-wise prediction problem, where the depth of each pixel is usually estimated independently. However, artifacts are often observed in the estimated depth map, e.g.*, depth values for points located in the same region may jump dramatically. To address this issue, we propose a novel self-supervised monocular depth estimation framework called GeoDepth, where we explore the intrinsic geometric representation in 3D scenes for producing accurate and continuous depth maps. In particular, we model the complex 3D scene as a collection of planes with varying sizes, where each plane is characterized by a unique set of parameters, namely planar normal (indicating plane orientation) and planar offset (defining the perpendicular distance from the camera center to the plane). Under this modeling, points in the same plane are enforced to share a unique representation and their depth variations related only to pixel coordinates, thus this geometric relationship can be exploited to regularize the depth variations of these points. To this end, we design a structured plane generation module that introduces spatio-temporal geometric cues and the plane uniqueness principle to recover the correct scene plane representation. In addition, we develop a depth discontinuity module to identify depth discontinuity regions and subsequently optimize them. Our experiments on the KITTI and NYUv2 datasets demonstrate that GeoDepth achieves state-of-the-art performance, with additional tests on Make3D and ScanNet validating its generalization capabilities.*

1. Introduction

Monocular depth estimation is a pivotal task in computer vision, which aims to predict a pixel-wise depth map for a given image. Each value in the depth map measures the distance from the corresponding point in the 3D scene to the camera plane [44]. Given its ability to provide 3D scene structure information, monocular depth estimation is broadly applicable in fields such as autonomous driving [11, 33], robotics [7], and augmented reality [31]. While supervised learning can deliver high-performance depth predictions, acquiring dense depth labels is often costly. As a result, self-supervised monocular depth estimation methods, which leverage only video or stereo images for training [9, 13, 46, 60], have emerged as a practical alternative.

Despite its advantages, self-supervised monocular depth estimation frequently encounters difficulties in achieving accurate and smooth depth estimation, particularly in regions lacking texture [27, 55, 57]. This mainly because the mainstream approaches usually adopts a ***point-to-depth*** modeling strategy (refer to Fig. 1(a)), where image pixels are treated as independent points, and the depth values are estimated from their discrete representations[28]. As a result, artifacts are often observed in the estimated depth map, *e.g.*, depth values for points located in the same region may jump dramatically. We refer to this phenomenon as depth discontinuity. Some works [27, 36, 54, 55, 57, 59] attempt to leverage only plane cues from surface normals to facilitate smoother depth estimation by assuming neighboring pixels form a local plane. Nevertheless, relying solely on surface normals as plane indicators might be unreliable, as it contradicts the principle of plane uniqueness in Geometric Algebra [20], which necessitates both a normal vector and an offset for a definitive plane representation in 3D space. Therefore, it is appealing to develop a robust modeling strategy to capture the intrinsic geometric representation for achieving accurate and continuous depth estimation.

In this paper, we propose a new self-supervised monocular depth estimation framework called GeoDepth, where a novel ***plane-to-depth*** modeling strategy is developed to address the depth discontinuity issues inherent in point-to-depth methods. The motivation of GeoDepth is that a 3D scene can essentially be modeled as a collection of planes of varying sizes, where all points within the same plane can be described by a unique set of parameters, namely the planar normal and planar offset [20]. As depicted in Fig. 1(b), our methodology models coplanar points as a uniform pla-

*Corresponding author

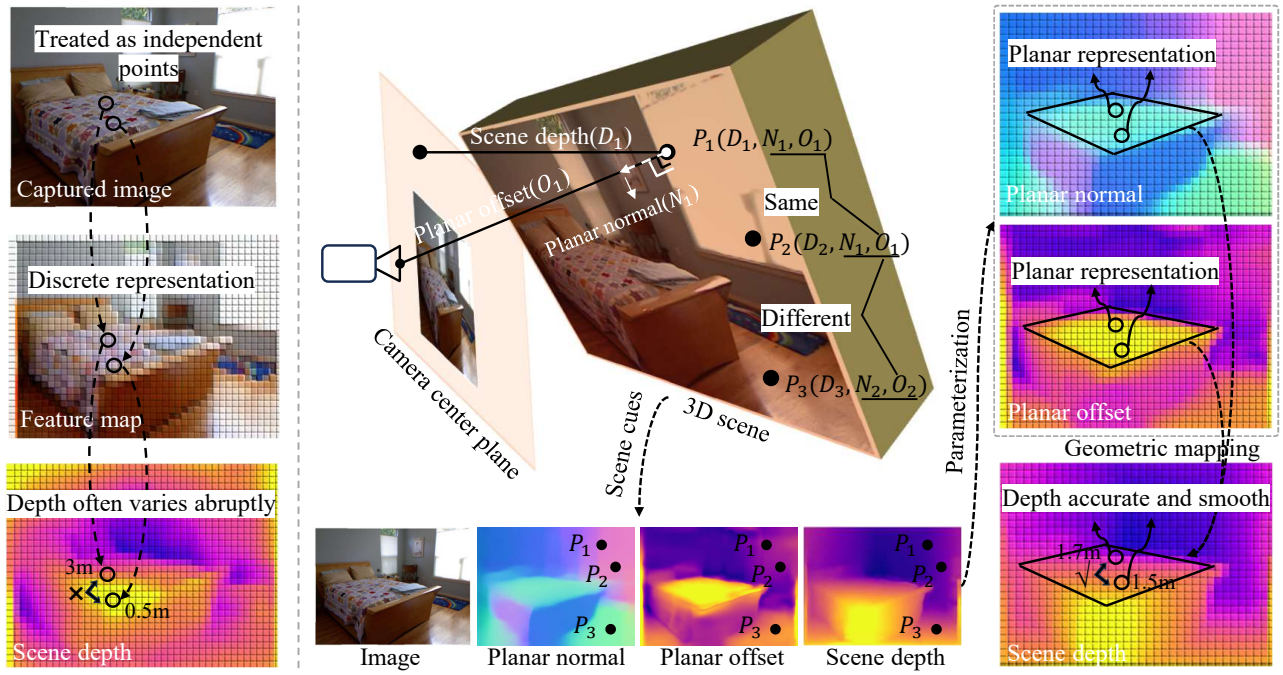


Figure 1. **An illustration of our motivation**. (a) Self-supervised monocular depth estimation has long been treated as a point-wise prediction problem, where the depth of each pixel is usually estimated independently. Consequently, artifacts are often observed in the estimated depth map, *e.g.*, depth values for points located in the same region may jump dramatically. (b) Instead, our GeoDepth models the complex 3D scene as planes of varying size, parameterized by planar normals and planar offsets, and infers the geometric correlation from them to depth. In GeoDepth, coplanar points are modeled as a unified representation instead of discrete representation and their depth variations are solely related to continuous image coordinates (see Eq. (1)). Therefore, by recovering structured representation among pixels, we are able to obtain more continuous depth map.

nar representation, rather than discrete points. This insight naturally leads us to estimate these two planar properties from the input image as a structured plane representation and derive the geometric correlation to depth from them. In our modeling framework, the depth variations of coplanar points are exclusively linked to pixel coordinates (see Eq. (1)), thereby eliminating depth discontinuities by restoring the structured representation among pixels.

To accurately recover the plane representation, we propose a structured plane generation module. This module initially utilizes spatio-temporal geometric cues to guide the two planar properties to recover the approximate scene structure, and then jointly optimizes them based on the plane uniqueness principle [20] to ensure that coplanar points converge to a unified representation. Additionally, recognizing that depth discontinuities often occur in low-texture regions, we introduce a depth discontinuity awareness module that dynamically identifies pixels with lower photometric errors as potential depth discontinuity regions and subsequently optimizes them.

We conduct extensive experiments on two outdoor datasets: KITTI [11] and Make3D [37], and two indoor datasets: NYUv2 [42] and ScanNet [6]. Results on KITTI and NYUv2 show that our method achieves state-of-the-art performance and that our design choices are effective. Results on Make3D and ScanNet verify the generalization ability of our method to unseen outdoor and indoor scenes.

In summary, our main contributions are as follows:

- We introduce GeoDepth, a novel self-supervised monocular depth estimation framework, which develops a ***plane-to-depth*** modeling strategy to address the depth discontinuity issues inherent in ***point-to-depth*** methods.
- We propose a structured plane generation module and a depth discontinuity awareness module to learn correct plane representations (*i.e.*, normal and offset) in a self-supervised paradigm.
- Extensive experiments on widely-used outdoor datasets (KITTI and Make3D) and indoor datasets (NYUv2 and ScanNet) show the robustness and generalization ability.

2. Related Work

2.1. Supervised Monocular Depth Estimation

As a pioneering work on supervised learning, Eigen *et al.* [8] first predicted a depth from a single image by stacking coarse-scale and fine-scale networks. Since then, many su-

pervised works have been proposed by modifying the model architecture [38, 50, 56], training on large datasets [5, 51], designing robust loss functions [26, 61], and transforming this task into an ordinal regression issue [2, 3, 21, 39]. However, we can only capture sparse depth data from the physical world, and collecting large-scale training data is time-consuming and expensive. Thus, high data requirements and collection costs limit the application of these methods.

2.2. Self-supervised Monocular Depth Estimation

Considering that large-scale annotated datasets are not always available, self-supervised depth estimation methods have attracted some attention. Garg *et al.* [10] first introduced an alternative to learn depth from stereo pairs in a self-supervised manner. To release the requirement of stereo images, Zhou *et al.* [66] estimated depth map and camera pose simultaneously and only used video data in the training phase. Following this, a series of works were proposed by designing robust losses [16, 41, 58], using auxiliary information during training [25, 48], dealing with moving objects [13, 24], and adding extra geometric constraints [27, 55]. However, these methods adopt point-to-depth modeling, that is, independently modeling the depth of each pixel [28], leading to depth discontinuities. In contrast, we revisit this task and propose a novel self-supervised framework, which utilizes plane-to-depth modeling to tackle the depth discontinuities.

2.3. Plane Cues for Depth Estimation

The purpose of depth estimation is to obtain depth maps with continuous structure. To this end, one type of monocular depth estimation work attempts to introduce multi-plane representations. GonzalezBello *et al.* [14, 15] predefined multiple depth planes and then used a deep neural network to learn the predefined plane to which each pixel belongs. However, as real 3D scenes are complex and contain slanted planes, such representation is sensitive to discretization [47]. Another type of work focuses on incorporating local plane prior. Yang *et al.* [55] introduced an edge-aware depth-normal consistency constraint inside the network to regularize the depth map by using the surface normal extracted from the depth map. Subsequently, some works [36, 54, 59] jointly predicted the depth map and the surface normal, and regularized the depth with a depth-normal consistency loss. In addition, StructDepth [27] adopted a Manhattan-world assumption in indoor settings, forcing major surfaces (floor, ceiling, and walls) to align with dominant directions. However, these methods adopted solely surface normals as plane guides, which is unreliable since it cannot represent the scene plane deterministically. Moreover, as non-planar region violates the local plane assumption, planar guidance in these methods is performed on all pixels, leading to incorrect depth training.

3. Plane-to-Depth Modeling

In monocular depth estimation, the objective is to produce a depth map where each value corresponds to the distance between the camera plane and its corresponding point in the 3D scene. Consequently, once the 3D scene has been reconstructed, the depth map can be readily derived using the principles of perspective geometry.

Typically, a real-world 3D scene can be seen as an assembly of planes of varying dimensions, ranging from the entire scene down to a single pixel. According to the plane uniqueness representation principle, a plane in 3D space can be represented by a normal vector $\boldsymbol{n}$ and an offset vector $\boldsymbol{o}$. Thus, a complete 3D scene can be mathematically denoted as: $\pi = \{\boldsymbol{n}_k, \boldsymbol{o}_k\}_{k=1}^{M}$, where M is the number of planes in the 3D scene. According to the point-normal form, a 3D point $\boldsymbol{P}$ on a plane $\pi_i = \{\boldsymbol{n}_i, \boldsymbol{o}_i\}$ satisfies: $\boldsymbol{n}_i^T \boldsymbol{P} = \boldsymbol{o}_i$, In addition, the 3D point $\boldsymbol{P}$ located in the plane π_i can be projected to a 2D point $\boldsymbol{p} = [u, v]^T$ in the image space: $\widetilde{\boldsymbol{p}} = \boldsymbol{P}/(\boldsymbol{d}\boldsymbol{K}^{-1})$, where $\widetilde{\boldsymbol{p}}$ denotes the homogeneous coordinate of $\boldsymbol{p}$, $\boldsymbol{d}$ is the depth at $\boldsymbol{P}$, and $\boldsymbol{K}$ is intrinsic to the camera. Thus, for an individual plane, the geometric correlation between the depth and planar normal as well as planar offset is defined as:

$$\boldsymbol{d} = \frac{\boldsymbol{o}_i}{\boldsymbol{n}_i^T \boldsymbol{K}^{-1} \widetilde{\boldsymbol{p}}}. \tag{1}$$

It can be observed that the depth variations of coplanar points are exclusively linked to pixel coordinates $\boldsymbol{p}$, since their planar normal and planar offset are equal. This insight naturally leads us to estimate these two planar properties from the input image as a structured plane representation and derive the geometric correlation to depth from them. To this end, we extend Eq. (1) to formulate the plane-to-depth modeling for self-supervised monocular depth estimation:

$$\boldsymbol{D}(\boldsymbol{p}) = \frac{\boldsymbol{O}(\boldsymbol{p})}{\boldsymbol{N}^T(\boldsymbol{p}) \boldsymbol{K}^{-1} \widetilde{\boldsymbol{p}}}, \tag{2}$$

where $\boldsymbol{N} \in \mathbb{R}^{3\times H\times W}$, $\boldsymbol{O} \in \mathbb{R}^{3\times H\times W}$ and $\boldsymbol{D} \in \mathbb{R}^{3\times H\times W}$ are the planar normal, planar offset, and depth map of an image, respectively. Note that $H \times W$ represent the spatial resolution of the input image.

Compared with the direct prediction of pixel-wise depth in existing self-supervised monocular depth estimation framework, our approach focuses on mining structured explicit plane representation and enjoys the following benefits. First, under our plane-to-depth modeling, coplanar points are modeled as a unified representation instead of discrete representation and their depth variations are solely related to continuous image coordinates. Thereby, we can avoid depth discontinuities by recover the correct structured scene representation. Second, estimating depth by fitting a plane effectively smooth local noise and reduce errors due to missing texture or other interference factors.

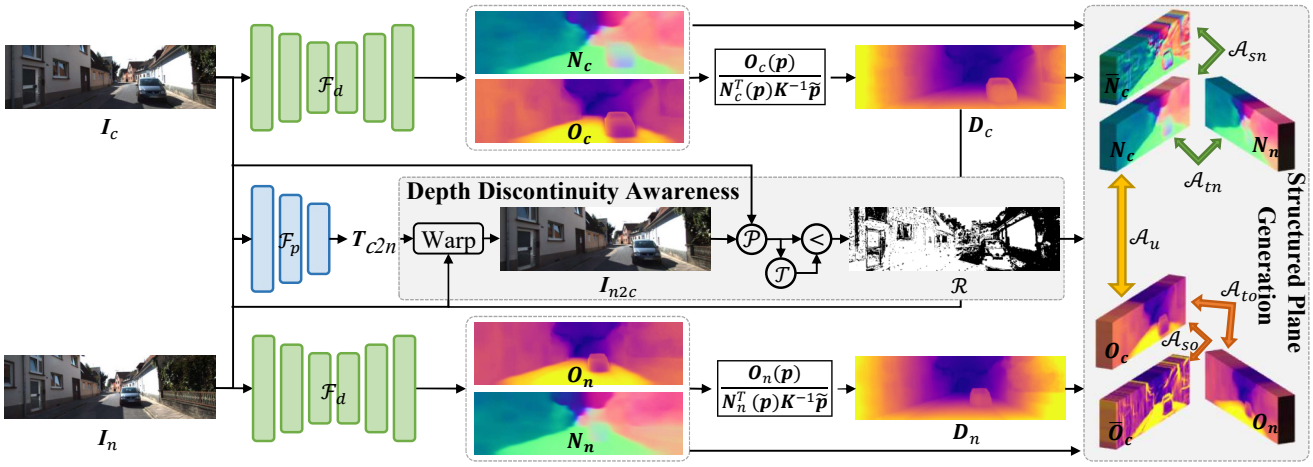


Figure 2. **Overview of the proposed method**. Instead of estimating the depth map from the input image, we first predict the planar normal and planar offset as structured plane representation, where the discontinuous regions will be detected by our proposed depth discontinuity awareness module and corrected by our proposed structured plane generation, from which the depth map will then be inferred.

4. The GeoDepth Model

Based on our plane-to-depth modeling described in Section 3, we propose GeoDepth to address the depth discontinuity issues inherent in point-to-depth methods. The overview of our proposed GeoDepth is presented in Fig. 2. In our framework, instead of estimating the pixel-wise depth map from the input image, we first predict the planar normal and planar offset as structured plane representation, where the discontinuous regions will be detected by our proposed depth discontinuity awareness module and corrected by our proposed structured plane generation, from which the depth map will then be inferred. Benefiting from our plane-to-depth modeling and corresponding plane constraints, coplanar points are modeled as a unified representation instead of discrete representation and their depth variations are solely related to continuous image coordinates. This results in an accurate and continuous depth map.

In the following sections, we first detail the two key components: structured plane generation module (Sec. 4.1) and depth discontinuity awareness module (Sec. 4.2). Then, we provide our self-supervised training scheme (Sec. 4.3).

4.1. Structured Plane Generation

To recover the correct plane representation, we propose a structured plane generation (SPG) module, which is depicted in Fig. 2. See supplementary materials for details. This module first uses spatio-temporal geometric cues to guide two planar properties to recover the approximate scene structure, and then jointly optimize them to ensure that coplanar points recover a unified representation.

Planar Normal Alignment. We exploit spatial and temporal geometric cues to align the estimated normals. Specifically, we first restore the scene point cloud using the depth map by: $\boldsymbol{P} = \boldsymbol{D}\boldsymbol{K}^{-1}\widetilde{\boldsymbol{p}}$, and select eight neighbor points to infer the normal of the central point, as shown in Fig. 3. Then, based on the orthogonality, the normal can be calculated by cross-product: $\widetilde{\boldsymbol{n}}_z = \overrightarrow{\boldsymbol{P}_z\boldsymbol{P}_{zx}} \times \overrightarrow{\boldsymbol{P}_z\boldsymbol{P}_{zy}}$, where $\boldsymbol{P}_z$ is the central point, and $\boldsymbol{P}_{zx}$, $\boldsymbol{P}_{zy}$ are the neighbor point. To improve the robustness of the final derived normals, we compute the average of the four initial normals:

$$\bar{\boldsymbol{N}}_c = \frac{1}{4}\sum_{z=1}^{4}(\widetilde{\boldsymbol{n}}_z). \tag{3}$$

Using the derived normals, the normal alignment in the spatial dimension is defined as:

$$\mathcal{A}_{sn} = 1 - cos(\Psi^{1/\sigma}(\bar{\boldsymbol{N}}_c), \Psi^{1/\sigma}(\boldsymbol{N}_c)). \tag{4}$$

where $\Psi^{1/\sigma}$ is the interpolation function with a downsampling rate of $1/\sigma$. We introduce downsampling to alleviate the impact of boundary pixels.

Additionally, to ensure the temporal consistency of planar geometry, we further explore temporal normal alignment. Given the planar normal $\boldsymbol{N}_n$ of the neighboring view, we can use the camera pose $\boldsymbol{T}_{c2n} = [\boldsymbol{R}_{c2n}|\boldsymbol{t}_{c2n}]$ to infer its observation normal in the target view: $\boldsymbol{N}'_n = \boldsymbol{R}_{n2c}\boldsymbol{N}_n$, where $\boldsymbol{R}_{n2c}$ is the inverse of rotation matrix $\boldsymbol{R}_{c2n}$. Subsequently, the planar normal of the target frame can be recovered by warping the observation normal $\boldsymbol{N}'_n$:

$$\boldsymbol{N}_{n2c} = \boldsymbol{N}'_n\langle proj(\boldsymbol{D}_c, \boldsymbol{T}_{c2n}, \boldsymbol{K})\rangle \tag{5}$$

After obtaining the transformed planar normal, the normal alignment in the temporal dimension is defined as:

$$\mathcal{A}_{tn} = \min_n(1 - cos(\Psi^{1/\sigma}(\boldsymbol{N}_{n2c}), \Psi^{1/\sigma}(\boldsymbol{N}_c))). \tag{6}$$

Using Eq. (6), we not only foster geometric consistency among samples within a batch but also extend this consistency across the entire video sequence.

Combining Eq. (4) and Eq. (6), the complete normal alignment can be defined as:

$$\mathcal{A}_n = \mathcal{A}_{sn} + \mathcal{A}_{tn}. \tag{7}$$

Planar Offset Alignment. We also mine the spatial and

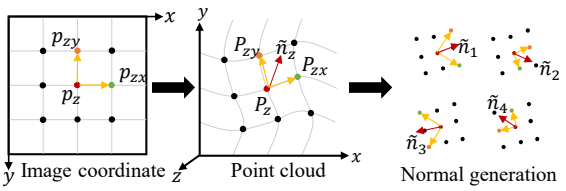


Figure 3. The illustration of Depth-to-Normal.

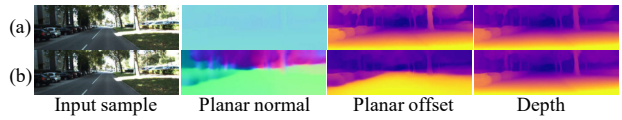


Figure 4. Visualization of scene geometric cues. (a) Without SPG, the plane representation is inaccurate. (b) With SPG guidance, the plane representations are consistent, and the depth map is accurate.

temporal geometric cues to align planar offset. According to the Eq. (2), given the derived planar normal $\bar{\boldsymbol{N}}_c$ and estimated depth $\boldsymbol{D}_c$, we can infer a pseudo planar offset:

$$\bar{\boldsymbol{O}}_c(\boldsymbol{p}) = \boldsymbol{D}_c(\boldsymbol{p})\bar{\boldsymbol{N}}_c^T(\boldsymbol{p})\mathbf{K}^{-1}\widetilde{\boldsymbol{p}} \tag{8}$$

Thus, the spatial offset alignment is defined as:

$$\mathcal{A}_{so} = \mathcal{P}(\Psi^{1/\sigma}(\boldsymbol{O}_c), \Psi^{1/\sigma}(\bar{\boldsymbol{O}}_c)), \tag{9}$$

where $\mathcal{P}$ is the photometric error [13].

Then, we mine temporal cues to align the estimate planar offset. Given the depth map $\boldsymbol{D}_c$ of captured image, we can obtain its observation depth $\boldsymbol{D}'_n$ in the neighboring view based on the warp operation, and then infer the observation planar offset of the neighboring view:

$$\mathbf{O}'_n(p) = \boldsymbol{D}'_n\boldsymbol{N}_n^T(p)\mathbf{K}^{-1}\widetilde{\boldsymbol{p}} \tag{10}$$

Besides, the sample planar offset of source view can be calculated by interpolating estimated offset $\boldsymbol{O}_n$:

$$\boldsymbol{O}''_n = \boldsymbol{O}_n \left\langle proj(\boldsymbol{D}_c, \boldsymbol{T}_{c2n}, \boldsymbol{K}) \right\rangle \tag{11}$$

It is noted that we cannot directly use $\boldsymbol{O}_n$ because the warping flow does not lie on the pixel grid. Finally, the temporal planar offset alignment is defined as:

$$\mathcal{A}_{to} = \min_n \mathcal{P}(\Psi^{1/\sigma}(\boldsymbol{O}'_n), \Psi^{1/\sigma}(\boldsymbol{O}''_n)). \tag{12}$$

Similar to the temporal normal alignment, this constraint allows the consistency to be propagated to the entire video sequence during the training stage.

Similar to planar normal alignment, our spatio-temporal offset alignment can be defined as:

$$\mathcal{A}_o = \mathcal{A}_{so} + \mathcal{A}_{to}. \tag{13}$$

Planar Uniqueness Alignment. By using the planar normal alignment and the planar offset alignment respectively, we can recover the roughly planar structure of the two planar properties. However, these two planar properties do not exist in isolation. According to the plane uniqueness representation principle, all points in a same plane can be represented by sharing a unique set of planar normal and planar offset. To this end, we jointly optimize the planar normal and planar offset by penalizing their first-order gradients to ensure that coplanar points recover a unified representation:

$$\mathcal{A}_u = \left|\nabla\Psi^{1/\sigma}(\boldsymbol{N}_c)\right| + \left|\nabla\Psi^{1/\sigma}(\boldsymbol{O}_c)\right|. \tag{14}$$

Eq.(14) enforces consistent gradient variations, reinforcing the geometric consistency of both attributes so that they collectively represent the same plane across coplanar points.

The impact of SPG on scene geometry estimation is shown Fig. 4. Without SPG, the plane representation is inaccurate, while with SPG guidance, the plane representations are consistent, and the depth map is accurate.

4.2. Depth Discontinuity Awareness

We know that depth discontinuity typically manifest in low-texture regions, where photometric errors are usually small and insufficient to provide effective supervision. To localize these discontinuity regions, we propose a depth discontinuity awareness module, which adaptively selects areas with lower photometric errors. This module is illustrated in Fig. 2. Using the inferred depth map $\boldsymbol{D}$, the predicted camera pose $\boldsymbol{T}_{c2n}$, and the neighboring image $\boldsymbol{I}_s$, we first synthesize the target image $\boldsymbol{I}_{n2c}$. Then, we calculate the per-pixel photometric errors $\mathcal{P}(\boldsymbol{I}_c, \boldsymbol{I}_{n2c})$. Finally, discontinuous regions can be identified by a threshold $\mathcal{T}$:

$$\mathcal{R} = \left[\mathcal{P}(\boldsymbol{I}_c, \boldsymbol{I}_{n2c}) < \mathcal{T}\right], \tag{15}$$

where $[\cdot]$ is the Iverson bracket. For $\mathcal{R}$, a value of 1 corresponds to discontinuous depth, and 0 to continuous depth.

Obviously, it is unreasonable to use a fixed threshold due to the varying degree of low-frequency components in each image. For this, we propose to use an adaptive threshold. Considering that the photometric error will gradually decrease with the improvement of depth accuracy during training, $\mathcal{T}$ is designed to be the mean of the photometric errors:

$$\mathcal{T} = \frac{1}{H \times W}\sum_p \mathcal{P}(\boldsymbol{I}_c, \boldsymbol{I}_{n2c}). \tag{16}$$

Besides, pixels in high-texture areas can usually estimate accurate depth and thus exhibit lower photometric errors, but will have slightly larger errors than low-texture pixels due to their own pixel intensity and bilinear interpolation. Thus, to filter out these pixels, we reformulate Eq. (15) as:

$$\mathcal{R} = \left[\mathcal{P}(\boldsymbol{I}_c, \boldsymbol{I}_{n2c}) < \mu \cdot \mathcal{T}\right], \tag{17}$$

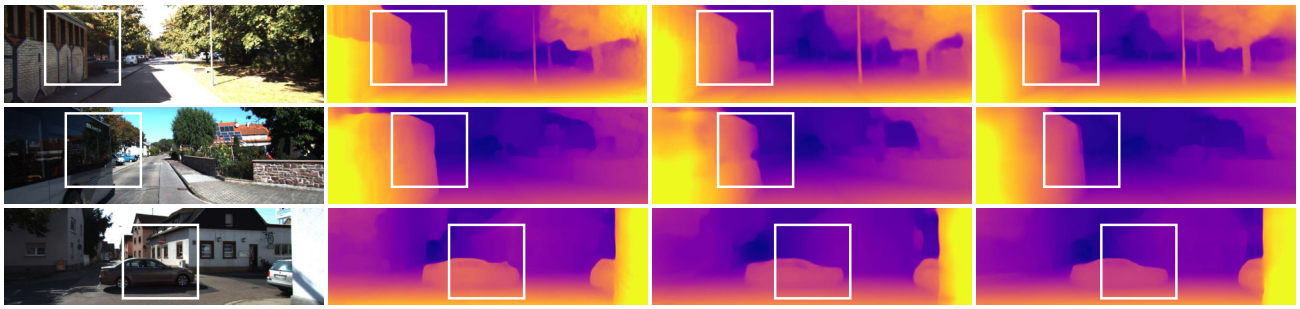
where μ is a balancing coefficient used to preserve discontinuity pixels. Using Eq. 17, we are able to accurately aware discontinuous regions and then avoid depth discontinuities by incorporating the structured plane generation module.

4.3. Self-Supervised Training

Similar to general self-supervised frameworks [13, 46, 60], our proposed GeoDepth consists of a depth network $\mathcal{F}_d$ and a pose network $\mathcal{F}_p$. Given a training instance $\{\boldsymbol{I}_c, \boldsymbol{I}_n\}$, we first use the $\mathcal{F}_d$ to estimate their two planar properties $\{\boldsymbol{N}_c, \boldsymbol{O}_c\}$ and $\{\boldsymbol{N}_n, \boldsymbol{O}_n\}$, respectively, and predict the relative pose $\boldsymbol{T}_{c2n}$ between them by $\mathcal{F}_p$. Subsequently, we infer the final depth according to Eq. (2). To learn a robust depth estimation model, in addition to the general loss $\mathcal{L}_g$, we add

Dataset	Method	Size	Mode	Train	Test	RMSE↓	RMSE log↓	Sq Rel↓	Abs Rel↓	δ<1.25↑	δ<1.25^2↑	δ<1.25^3↑
KITTI	Monodepth [12]	512×256	S	✓	✓	5.927	0.247	1.344	0.148	0.803	0.922	0.964
	3Net [35]	512×256	S	✓	✓	5.888	0.208	1.201	0.119	0.844	0.941	0.978
	Monodepth2 [13]	640×192	S	✓	✓	4.960	0.208	0.873	0.109	0.864	0.948	0.975
	BRNet [19]	640×192	S	✓	✓	4.716	0.197	0.876	0.103	0.954	0.978	
	Monodepth2 [13]	640×192	MS	✓	✓	4.750	0.196	0.818	0.106	0.874	0.957	0.979
	DepthHints [48]	640×192	MS	✓	✓	4.627	0.189	0.769	0.105	0.875	0.959	0.982
	HR-Depth [32]	640×192	MS	✓	✓	4.612	0.185	0.785	0.107	0.887	0.962	0.982
	R-MSFM6 [67]	640×192	MS	✓	✓	4.625	0.189	0.787	0.111	0.882	0.961	0.981
	Monodepth2 [13]	640×192	M	✓	✓	4.863	0.193	0.903	0.115	0.877	0.959	0.971
	HR-Depth [32]	640×192	M	✓	✓	4.632	0.185	0.792	0.109	0.884	0.962	0.983
	CADepth-Net [53]	640×192	M	✓	✓	4.535	0.181	0.769	0.105	0.892	0.964	0.983
	DIFFNet [64]	640×192	M	✓	✓	4.483	0.180	0.764	0.102	0.896	0.965	0.983
	MonoFormer [1]	640×192	M	✓	✓	4.580	0.183	0.846	0.104	0.891	0.962	0.982
	SC-DepthV3 [45]	640×192	M	✓	✓	4.709	0.188	0.756	0.118	0.864	0.960	0.984
	SRD-Depth [30]	640×192	M	✓	✓	4.619	0.186	0.762	0.111	0.877	0.961	0.983
	Swin-Depth [40]	640×192	M	✓	✓	4.510	0.182	0.739	0.106	0.890	0.964	0.984
	Lite-Mono [60]	640×192	M	✓	✓	4.561	0.183	0.765	0.107	0.886	0.963	0.983
	ShuffleMono [29]	640×192	M	✓	✓	4.821	0.193	0.850	0.114	0.872	0.957	0.980
	Liu *et al.* [29]	640×192	M	✓	✓	4.724	0.187	0.747	0.114	0.863	0.960	0.984
	Dynamo-Depth [46]	640×192	M	✓	✓	4.505	0.183	0.758	0.112	0.873	0.959	0.984
	GeoDepth	640×192	M	✓	✓	**4.381**	**0.176**	**0.694**	**0.100**	**0.897**	**0.966**	**0.984**
Make3D	Monodepth2 [13]	640×192	M	×	✓	7.418	0.163	3.589	0.322	-	-	-
	HR-Depth [32]	640×192	M	×	✓	7.024	0.159	3.208	0.315	-	-	-
	CADepth-Net [53]	640×192	M	×	✓	7.066	0.159	3.086	0.312	-	-	-
	DIFFNet [64]	640×192	M	×	✓	7.008	0.155	3.313	0.309	-	-	-
	Lite-Mono [60]	640×192	M	×	✓	6.981	0.158	3.060	0.305	-	-	-
	Zhao *et al.* [62]	640×192	M	×	✓	7.095	0.158	3.200	0.316	-	-	-
	Xiong *et al.* [52]	640×192	M	×	✓	7.005	0.161	3.102	0.319	-	-	-
	GeoDepth	640×192	M	×	✓	**6.735**	**0.153**	**2.750**	**0.296**	-	-	-

Table 1. **Quantitative depth estimation results on the KITTI** [11] and **Make3D** [37] **datasets**. Comparison methods are grouped by training mode: S indicates stereo training, M indicates monocular training, and MS indicates both stereo and monocular training. The KITTI dataset is used to assess within-dataset performance, while the Make3D dataset is used to evaluate cross-dataset generalization. "✓" in the train column or test column means training or testing in this dataset, while "×" in the train column means not training in this dataset. "-" indicates not applicable. The best results are in **bold**.



(a) Input (b) Monodepth2 (c) Lite-Mono (d) Ours

Figure 5. Qualitative results on the KITTI dataset [11]. The depth maps predicted by MonoDepth2 and Lite-Mono exhibit inconsistencies in planar regions and noticeable errors along object edges. In contrast, our method produces smoother and more accurate depth estimations, effectively preserving both planar structures and sharp boundaries.

a structured planar generation loss $\mathcal{L}_{sp}$ during training. Our overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_{sp}. \tag{18}$$

The $\mathcal{L}_g$ consists of a minimum photometric loss and an edge-ware smoothness loss: $\mathcal{L}_g = \mathcal{L}_{ph} + \lambda\mathcal{L}_{sm}$, where $\mathcal{L}_{ph} = \sum \min_n \mathcal{P}(\boldsymbol{I}_c, \boldsymbol{I}_{n2c})$, $\mathcal{L}_{sm} = \sum |\partial_x \boldsymbol{D}_c^*| \, e^{-|\partial_x \boldsymbol{I}_c|} + |\partial_y \boldsymbol{D}_c^*| \, e^{-|\partial_y \boldsymbol{I}_c|}$ and λ is fixed to 0.001. $\boldsymbol{D}_c^* = \boldsymbol{D}_c / \bar{\boldsymbol{D}}_c$ is the mean-normalized inverse depth. $\mathcal{L}_{sp}$ is defined as: $\mathcal{L}_{sp} = \sum \Psi^{1/\sigma}(\mathcal{R})(\gamma\mathcal{A}_n + \varphi\mathcal{A}_o + \omega\mathcal{A}_u)$, where γ, φ, and ω are set as 0.03, 0.01 and 0.1.

Dataset	Method	Size	Mode	Train	Test	RMSE↓	Abs Rel↓	δ<1.25↑	δ<1.25^2↑	δ<1.25^3↑
NYUv2	MovingIndoor [65]	320×256	M	✓	✓	0.712	0.208	0.674	0.900	0.968
	Monodepth2 [13]	320×256	M	✓	✓	0.601	0.160	0.767	0.949	0.988
	P^{2}Net [57]	320×256	M	✓	✓	0.599	0.159	0.772	0.942	0.984
	SC-DepthV1 [4]	320×256	M	✓	✓	0.639	0.159	0.734	0.937	0.983
	PLNet [22]	320×256	M	✓	✓	0.562	0.151	0.790	0.953	0.989
	StructDepth [27]	320×256	M	✓	✓	0.540	0.142	0.813	0.954	0.988
	ADPDepth [43]	320×256	M	✓	✓	0.592	0.165	0.753	0.934	0.981
	F^{2}Depth [17]	320×256	M	✓	✓	0.569	0.153	0.787	0.950	0.987
	Guo *et al.* [18]	320×256	M	✓	✓	0.567	0.152	0.792	0.950	0.988
	Ours	320×256	M	✓	✓	**0.520**	**0.134**	**0.833**	**0.963**	**0.991**
ScanNet	MovingIndoor [65]	320×256	M	×	✓	0.483	0.212	0.650	0.905	0.976
	Monodepth2 [13]	320×256	M	×	✓	0.458	0.200	0.672	0.922	0.981
	TrainFlow [63]	320×256	M	×	✓	0.415	0.179	0.726	0.927	0.980
	P^{2}Net [57]	320×256	M	×	✓	0.420	0.175	0.740	0.932	0.982
	PLNet [22]	320×256	M	×	✓	0.414	0.176	0.735	0.939	0.985
	IFMNet [49]	320×256	M	×	✓	0.402	0.170	0.758	0.940	0.989
	SC-Depthv1 [4]	320×256	M	×	✓	0.392	0.169	0.749	0.938	0.983
	StructDepth [27]	320×256	M	×	✓	0.400	0.165	0.754	0.939	0.985
	GeoDepth	320×256	M	×	✓	**0.387**	**0.161**	**0.769**	**0.946**	**0.987**

Table 2. **Quantitative depth estimation results on the NYUv2 [42] and ScanNet [6] datasets**. The NYUv2 dataset is used to assess within-dataset performance, while the ScanNet dataset is used to evaluate cross-dataset generalization.

5. Experiments

5.1. Datasets and Evaluation Metrics

Outdoor Datasets. **KITTI** [11] is a standard outdoor benchmark. Following Monodepth2 [13], we use the data split of Eigen *et al.* [8], which has 39,180 monocular triplets for training, 4,424 for evaluation, and 697 for testing. **Make3D** [37] is often used as outdoor generalization test. It consists 134 images. The image is resized to 640×192 during both training and testing.

Indoor Datasets. **NYUv2** [42] contains 464 videos captured by a hand-held camera. Following StructDepth [22], we use 654 images for testing and the remaining 335 sequences for training and validation. **ScanNet** [6] provides 1513 indoor videos, captured by handheld devices. We use the test split in StructDepth [27], including 533 images, to verify generalization. The image is resized to 320×256 during both training and testing.

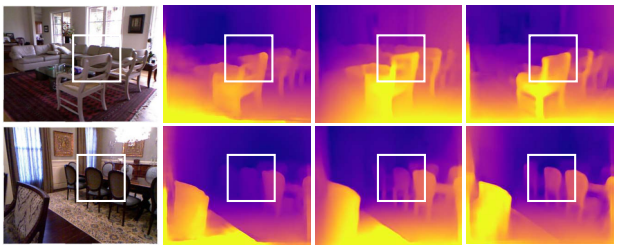
Evaluation Metrics. We follow Monodepth2 [13] using absolute relative error (**Abs Rel**), relative squared error (**Sq Rel**), root mean squared error (**RMSE**), root mean squared logarithmic error (**RMSE log**), and threshold accuracy (σ<1.25, σ<1.25^2 and σ<1.25^3).

5.2. Implementation Details

We implemented model in PyTorch [34] using an NVIDIA RTX 3090 GPU. Our models are trained for 20 epochs on the outdoor dataset and 40 epochs on the indoor dataset using Adam optimizer [23], with a batch size of 12 and 16, respectively. The learning rate is set to 10^{-4}, and decays to 10^{-5} after 15 epoch for outdoor dataset and after 20 epoch for indoor dataset. The α, σ, and μ are set to 0.85, 2, and 0.5, respectively. Our goal was to validate the core idea of plane-to-depth modeling, so we did not incorporate performance-improving techniques from recent SOTA methods (*e.g.*, arbitrary-scale data augmentation in RA-Depth and the transformer-based feature extractor in MonoViT). Instead, we opted for a simpler architecture (only borrowing the DepthNet and PoseNet from RA-Depth) to avoid introducing excessive complexity and to focus on demonstrating the effectiveness of our method. Thus, we did not include a direct comparison with these works. However, the results of integrating our idea with SOTA methods are reported in the supplementary material.

5.3. Comparison on Outdoor Scene

KITTI. Table 1 provides quantitative results of training and testing on the outdoor dataset, i.e., KITTI dataset. Our method outperforms existing self-supervised approaches on all metrics. In particular, our method relatively outperforms the classical method Swin-Depth by 6.1% and by 5.7% in terms of Sq Rel and Abs Rel, respectively. Moreover, our method achieves superior performance when trained only on monocular videos, even compared to methods trained on stereo pairs or a combination of stereo pairs and monocular videos. We also compare the qualitative performance with the classic work Monodepth2 [13] and the recent work Lite-Mono [60]. As can be seen in Fig. 5, the depth maps predicted by MonoDepth2 and Lite-Mono are inconsistent in plane areas and exhibit noticeable errors at edges. In con-



(a) Input (b) StructDepth (c) PLNet (d) Ours

Figure 6. Qualitative results on the NYUv2 dataset [42].

trast, the depth generated by our method is smooth and accurate for both plane regions and edges.

Generalization Study on Make3D. To show the generalization capability in outdoor scenes, we test on Make3D [37]. All models were trained only on KITTI. Table 1 shows the quantitative comparison. Our method outperforms all previous work, indicating our method can be well generalized to unseen outdoor scenes.

5.4. Comparison on Indoor Scene

NYUv2. Compared to outdoor scenes, indoor scenes is challenging because it contains many low-texture planes. To verify the effectiveness in difficult scene, we trained and tested on NYUv2 [42], and the results are shown in Table 2. Our method outperforms all other self-supervised methods, which demonstrates that utilizing plane-to-depth modeling can improve the depth quality. Moreover, we provide the qualitative performance with self-supervised methods including focusing on indoor scenes, as shown in Fig. 6. It can be found that our proposed method obtain continuous depth results in region lacking texture.

Generalization Study on Make3D. Table 2 shows the generalization comparison of our method with state-of-the-art methods on ScanNet [6]. All models are trained on NYUv2 [42]. Our method outperforms all previous approaches, which demonstrates our method can be well generalized to unseen indoor scenes.

5.5. Ablation Study.

In Table 3, we present an ablation study evaluating the impact of three key components in our method: Plane-to-Depth Modeling (P2D), Depth Discontinuity Awareness (DDA), and Structured Plane Generation (SPG). These components are progressively added to a baseline model that predicts depth directly from a single image.

Effects of Plane-to-Depth Modeling. Incorporating the Plane-to-Depth modeling shows a notable improvement over the baseline. By modeling planes explicitly and inferring the depth from plane properties, RMSE decreases from from 4.471 to 4.436, and Sq Rel decreases 0.751 to 0.740, indicating our plane-to-depth modeling can efficiently leverage structured scene cues across pixels for enhancing depth accuracy. Furthermore, compared with the baseline framework, there are only few parameter increases, indicating that the performance improvement mainly comes from modeling rather than by increasing model parameters.

Method	P2D	SPG	DDA	RMSE ↓	Sq Rel↓	δ<1.25↑	#Params
Baseline				4.471	0.751	0.895	9.98M
+P2D	✓			4.436	0.740	0.896	10.0M
+P2D+SPG	✓	✓		4.412	0.722	0.896	10.0M
GeoDepth	✓	✓	✓	**4.381**	**0.694**	**0.897**	10.0M

Table 3. **Ablation study**. Baseline: Predict depth directly from a single image. P2D: Using Plane-to-Depth modeling. DDA: Using Depth Discontinuity Awareness module SPG: Using Structured Plane Generation module.

Effects of Structured Plane Generation. When the SPG module is added to the P2D model, there is a further reduction in RMSE and Sq Rel, with values dropping to 4.412 and 0.722, respectively. Moreover, the SPG module does not add any new model parameters. This configuration demonstrates that SPG significantly enhances the structured prediction of depth planes, thereby optimizing the depth performance of the model.

Effects of Depth Discontinuity Awareness. Finally, adding the DDA module, as seen in the full model (GeoDepth), yields the best results. The complete model achieves an RMSE of 4.381 and Sq Rel of 0.694, with σ reaching 0.897. Moreover, the DDA module does not add any new model parameters. This improvement highlights the ability of the DDA module to accurately identify discontinuous regions and then avoid depth discontinuities by incorporating the structured plane generation module.

6. Conclusion

In this paper, we introduce a paradigm shift for continuous depth estimation by proposing GeoDepth, which utilizes plane-to-depth modeling to overcome the depth discontinuity of point-to-depth methods. Specifically, our GeoDepth models the complex 3D scene as planes of varying size, parameterized by planar normals and planar offsets, and infers the depth from them. Then, we propose a structured plane generation module and a depth discontinuity awareness module to learn correct plane representations (*i.e.*, normal and offset) in a self-supervised paradigm. Extensive experiments on widely-used outdoor datasets (KITTI and Make3D) and indoor datasets (NYUv2 and ScanNet) show the robustness and generalization ability.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62176047), the National Natural Science Foundation of China (No. 62476051), the Sichuan Natural Science Foundation (No. 2024NSFTD0041), and the Sichuan Science and Technology Program (No. 2021YFS0374).

References

[1] Jinwoo Bae, Sungho Moon, and Sunghoon Im. Deep digging into the generalization of self-supervised monocular depth estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 187–196, 2023. 6

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 3

[4] Jia-Wang Bian, Huangying Zhan, Naiyan Wang, Zhichao Li, Le Zhang, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised scale-consistent depth learning from video. *International Journal of Computer Vision (IJCV)*, 2021. 7

[5] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 679–688, 2020. 3

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 7, 8

[7] Gregory Dudek and Michael Jenkin. *Computational principles of mobile robotics*. Cambridge university press, 2024. 1

[8] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 2, 7

[9] Yingwei Feng, Zhiyong Hong, Liping Xiong, Zhiqiang Zeng, and Jingmin Li. Shufflemono: Rethinking lightweight network for self-supervised monocular depth estimation. *Journal of Artificial Intelligence and Soft Computing Research*, 14(3):191–205. 1

[10] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 740–756. Springer, 2016. 3

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 6, 7

[12] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017. 6

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1, 3, 5, 6, 7

[14] Juan Luis Gonzalez and Munchurl Kim. Plade-net: Towards pixel-level accuracy for self-supervised single-view depth estimation with neural positional encoding and distilled matting loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6851–6860, 2021. 3

[15] Juan Luis GonzalezBello and Munchurl Kim. Forget about the lidar: Self-supervised depth estimators with med probability volumes. *Advances in Neural Information Processing Systems*, 33:12626–12637, 2020. 3

[16] Ariel Gordon, Hanhan Li, Rico Jonschkowski, and Anelia Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8977–8986, 2019. 3

[17] Xiaotong Guo, Huijie Zhao, Shuwei Shao, Xudong Li, and Baochang Zhang. F^2depth: Self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis. *Engineering Applications of Artificial Intelligence*, 133:108391, 2024. 7

[18] Xiaotong Guo, Huijie Zhao, Shuwei Shao, Xudong Li, Baochang Zhang, and Na Li. Sim-multidepth: Self-supervised indoor monocular multi-frame depth estimation based on texture-aware masking. *Remote Sensing*, 16(12): 2221, 2024. 7

[19] Wencheng Han, Junbo Yin, Xiaogang Jin, Xiangdong Dai, and Jianbing Shen. Brnet: Exploring comprehensive features for monocular depth estimation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022. 6

[20] Dietmar Hildenbrand, Christian Perwass, Leo Dorst, and Daniel Fontijne. Geometric algebra and its application to computer graphics. In *Eurographics (Tutorials)*, 2004. 1, 2

[21] Dongting Hu, Liuhua Peng, Tingjin Chu, Xiaoxing Zhang, Yinian Mao, Howard Bondell, and Mingming Gong. Uncertainty quantification in depth estimation via constrained ordinal regression. In *European Conference on Computer Vision*, pages 237–256. Springer, 2022. 3

[22] Hualie Jiang, Laiyan Ding, Junjie Hu, and Rui Huang. Plnet: Plane and line priors for unsupervised indoor depth estimation. In *2021 International Conference on 3D Vision (3DV)*, pages 741–750. IEEE, 2021. 7

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[24] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 582–600. Springer, 2020. 3

[25] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European conference on computer vision (ECCV)*, pages 698–713, 2018. 3

[26] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 3

[27] Boying Li, Yuan Huang, Zeyu Liu, Danping Zou, and Wenxian Yu. Structdepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12663–12673, 2021. 1, 3, 7

[28] Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, and Luc Van Gool. Single image depth prediction made better: A multivariate gaussian take. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17346–17356, 2023. 1, 3

[29] Runze Liu, Dongchen Zhu, Guanghui Zhang, Yue Xu, Wenjun Shi, Xiaolin Zhang, Lei Wang, and Jiamao Li. Unsupervised monocular depth estimation based on hierarchical feature-guided diffusion. *arXiv preprint arXiv:2406.09782*, 2024. 6

[30] Zhong Liu, Ran Li, Shuwei Shao, Xingming Wu, and Weihai Chen. Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 6

[31] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM Transactions on Graphics (ToG)*, 39(4):71–1, 2020. 1

[32] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2294–2301, 2021. 6

[33] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3061–3070, 2015. 1

[34] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 7

[35] Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International conference on 3d vision (3DV)*, pages 324–333. IEEE, 2018. 6

[36] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 1, 3

[37] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008. 2, 6, 7, 8

[38] Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, and Zhengguo Li. Nddepth: Normal-distance assisted monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7931–7940, 2023. 3

[39] Shuwei Shao, Zhongcai Pei, Xingming Wu, Zhong Liu, Weihai Chen, and Zhengguo Li. Iebins: Iterative elastic bins for monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[40] Dongseok Shim and H Jin Kim. Swindepth: Unsupervised depth estimation using monocular sequences via swin transformer and densely cascaded network. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4983–4990. IEEE, 2023. 6

[41] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, pages 572–588. Springer, 2020. 3

[42] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2, 7, 8

[43] Xiaogang Song, Haoyue Hu, Li Liang, Weiwei Shi, Guo Xie, Xiaofeng Lu, and Xinhong Hei. Unsupervised monocular estimation of depth and visual odometry uusing attention and depth-pose consistency loss. *IEEE Transactions on Multimedia*, 2023. 7

[44] Jaime Spencer, Fabio Tosi, Matteo Poggi, Ripudaman Singh Arora, Chris Russell, Simon Hadfield, Richard Bowden, GuangYuan Zhou, ZhengXin Li, Qiang Rao, et al. The third monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–14, 2024. 1

[45] Libo Sun, Jia-Wang Bian, Huangying Zhan, Wei Yin, Ian Reid, and Chunhua Shen. Sc-depthv3: Robust self-supervised monocular depth estimation for dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 6

[46] Yihong Sun and Bharath Hariharan. Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 5, 6

[47] Ruoyu Wang, Zehao Yu, and Shenghua Gao. Planedepth: Self-supervised depth estimation via orthogonal planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21425–21434, 2023. 3

[48] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-supervised monocular depth hints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2162–2171, 2019. 3, 6

[49] Yi Wei, Hengkai Guo, Jiwen Lu, and Jie Zhou. Iterative feature matching for self-supervised indoor depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3839–3852, 2021. 7

[50] Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019. 3

[51] Ke Xian, Chunhua Shen, Zhiguo Cao, Hao Lu, Yang Xiao, Ruibo Li, and Zhenbo Luo. Monocular relative depth percep-

tion with web stereo data supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 311–320, 2018. 3

[52] Mingkang Xiong, Zhenghong Zhang, Jiyuan Liu, Tao Zhang, and Huilin Xiong. Monocular depth estimation using self-supervised learning with more effective geometric constraints. *Engineering Applications of Artificial Intelligence*, 128:107489, 2024. 6

[53] Jiaxing Yan, Hong Zhao, Penghui Bu, and YuSheng Jin. Channel-wise attention-based network for self-supervised monocular depth estimation. In *2021 International Conference on 3D vision (3DV)*, pages 464–473. IEEE, 2021. 6

[54] Zhenheng Yang, Peng Wang, Yang Wang, Wei Xu, and Ram Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 225–234, 2018. 1, 3

[55] Zhenheng Yang, Peng Wang, Wei Xu, Liang Zhao, and Ramakant Nevatia. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 1, 3

[56] Songsong Yu, Yifan Wang, Yunzhi Zhuge, Lijun Wang, and Huchuan Lu. Dme: Unveiling the bias for better generalized monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6817–6825, 2024. 3

[57] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *European Conference on Computer Vision*, pages 206–222. Springer, 2020. 1, 7

[58] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 340–349, 2018. 3

[59] Huangying Zhan, Chamara Saroj Weerasekera, Ravi Garg, and Ian Reid. Self-supervised learning for single view depth and surface normal estimation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4811–4817. IEEE, 2019. 1, 3

[60] Ning Zhang, Francesco Nex, George Vosselman, and Norman Kerle. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18537–18546, 2023. 1, 5, 6, 7

[61] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Zequn Jie, Xiang Li, and Jian Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 3

[62] Hailiang Zhao, Yongyi Kong, Chonghao Zhang, Haoji Zhang, and Jiansen Zhao. Learning effective geometry representation from videos for self-supervised monocular depth estimation. *ISPRS International Journal of Geo-Information*, 13(6):193, 2024. 6

[63] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9151–9161, 2020. 7

[64] Hang Zhou, David Greenwood, and Sarah Taylor. Self-supervised monocular depth estimation with internal feature fusion. *arXiv preprint arXiv:2110.09482*, 2021. 6

[65] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving indoor: Unsupervised video depth learning in challenging environments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8618–8627, 2019. 7

[66] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1851–1858, 2017. 3

[67] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-msfm: Recurrent multi-scale feature modulation for monocular depth estimating. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12777–12786, 2021. 6