CORE: Co-planarity Regularized Monocular Geometry Estimation with Weak Supervision

Yuguang Li[1], Kai Wang[1], Hui Li[1], Seon-Min Rhee[2], Seungju Han[2], Jihye Kim[2],
Min Yang[1], Ran Yang[1], Feng Zhu[1]
[1]Samsung R&D Institute China Xi'an (SRCX)
[2]Samsung Advanced Institute of Technology (SAIT), South Korea
yg.li, k001.wang, hui01.li, s.rhee, sj75.han, jihye32.kim,
min16.yang, ran01.yang, f15.zhu@samsung.com

Abstract

The ill-posed nature of monocular 3D geometry (depth map and surface normals) estimation makes it rely mostly on data-driven approaches such as Deep Neural Networks (DNN). However, data acquisition of surface normals, especially the reliable normals, is acknowledged difficult. Commonly, reconstruction of surface normals with high quality is heuristic and time-consuming. Such fact urges methodologies to minimize dependency on ground-truth normals when predicting 3D geometry. In this work, we devise CO-planarity REgularized (CORE) loss functions and Structure-Aware Normal Estimator (SANE). Without involving any knowledge of ground-truth normals, these two designs enable pixel-wise 3D geometry estimation weakly supervised by only ground-truth depth map. For CORE loss functions, the key idea is to exploit locally linear depth-normal orthogonality under spherical coordinates as pixel-level constraints, and utilize our designed Adaptive Polar Regularization (APR) to resolve underlying numerical degeneracies. Meanwhile, SANE easily establishes multi-task learning with CORE loss functions on both depth and surface normal estimation, leading to the whole performance leap. Extensive experiments present the effectiveness of our method on various DNN architectures and data benchmarks. The experimental results demonstrate that our depth estimation achieves the state-of-the-art performance across all metrics on indoor scenes and comparable performance on outdoor scenes. In addition, our surface normal estimation is overall superior.

1. Introduction

Depth and surface normals are essential elements of 3D geometry. With the assistance of surface normals, depth map is able to faithfully describe the characteristics of the 3D scenes [4], which benefits various 3D applications, *e.g.*,

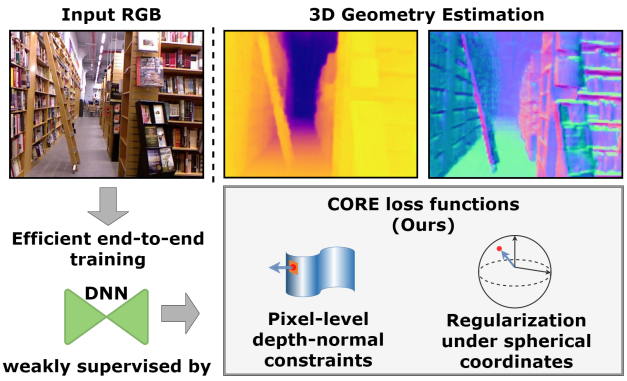


Figure 1: Our weakly supervised monocular 3D geometry estimation. CORE loss functions distinguish our method from others. Instead of proposing surface normal candidates in local or global depth regions, we properly regularize and utilize the pixel-level depth-normal constraints for pixel-wise loss functions, which enable the efficient end-to-end training with only ground-truth depth map.

3D reconstruction [33, 4], augmented reality [18] and autonomous driving [16]. To recover 3D geometry, monocular depth and surface normal estimation provide the most convenient solutions, and meanwhile define the challenging ill-posed problems.

Recent approaches mostly tackle these two ill-posed problems as genuine data-driven tasks by DNN regression or classification. However, it is widely acknowledged that data acquisition of surface normals is not a straightforward task [22, 8, 41, 3]. Particularly, the reliable surface normals are much more difficult to be obtained [22, 8, 3]. There are various commercial sensors such as LiDAR and ToF to acquire depth map directly, but no alternative available for surface normals. To acquire surface normals, the common practice is to involve least-square plane fitting from the depth maps [41]. Unfortunately, these depth maps cap-

tured by consumer-level sensors are usually contaminated by noises, resulting in deteriorated quality on generated surface normals [22, 41, 3]. As a compromise, heuristic and time-consuming post-processing is commonly involved during preparation of the ground truth normals [22, 8].

It is certainly a meaningful advantage to be able to predict surface normals by depending less or none of ground-truth normals. One idea is to employ pre-trained surface normal estimation network, followed by refinement that makes use of depth-normal consistency [38, 32, 34, 4]. On the condition of pre-training, the later refinement is effective even without ground-truth normals. However, the pre-training still involves the ground truth as prerequisites, and the depth-normal consistency is utilized for building refinement rather than guiding regression from scratch. Another idea is about the proposal of surface normal candidates from depth map or point cloud, including differentiable least-square [32, 30, 31], local differentiation [17, 20] and random sampling [41, 29]. Nevertheless, when the proposal of depth region shrinks to a small local area or finally pixel level, suboptimal and noisy surface normal candidates could appear. At this time, more supervision by ground-truth normals are demanded [17, 20, 29]. As for the global proposal, it would suffer from computational burden and insufficient local features [32, 30, 31, 41].

In this work, we propose novel Co-planarity Regularized (CORE) loss functions derived from regularized spherical depth-normal constraints at pixel level. These CORE loss functions establish pixel-wise surface normal regression weakly supervised by only ground-truth depth map. It is worth noting that we particularly express pixel-level depth-normal constraints with spherical coordinates, and regularize these constraints by the polar view. Similar to the occurrence of suboptimal surface normal candidates, depth-normal constraints are progressively weakened towards the pixel level, resulting in the emergence of degeneracies during back-propagation (see Sec. 3.1). By re-formulating depth-normal constraints under spherical coordinates, we observe that the degeneracies are mostly attributed to the polar angle collapse as shown in Fig. 2. To counter this, we devise a novel Adaptive Polar Regularization (APR) term as a part of CORE loss functions. This term adaptively penalizes polar angle estimation at the sub-optimal state, thus resolving the degeneracies. As a result, the regularized spherical depth-normal constraints can be used as stable pixel-level constraints that serve as loss functions to efficiently guide surface normal regression from scratch. Moreover, our surface normal prediction is visualized in Fig. 1 and Sec. 4, which presents plenty of local geometry details.

Another advantage of our method is that CORE loss functions are perfect for multi-task learning on both depth and surface normal estimation. Accordingly, we design a Structure-Aware Normal Estimator (SANE) that collaborates with our CORE loss functions. SANE can be easily applied as a side branch to existing encoder-decoder architectures of depth estimation as shown in Fig. 3. Driven by CORE loss functions, both depth estimation from the original architecture and surface normal estimation from SANE mutually achieves performance leap.

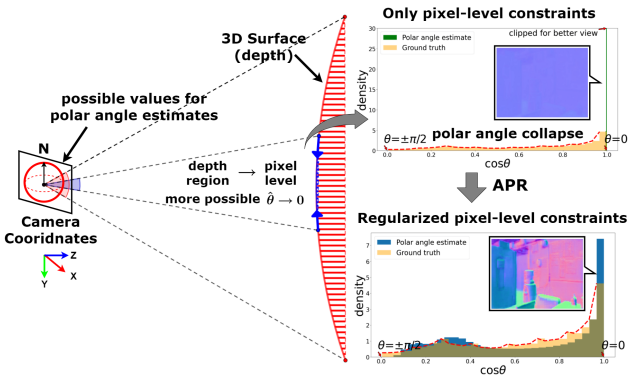


Figure 2: The polar angle collapse. When depth supervision is shrinking towards the pixel level, the depth-normal constraints are progressively weaken, which leads to the overall prediction distribution of polar angle $\hat{\theta}$ excessively concentrated near 0 (the green histogram). The polar angle estimates collapse in ranges other than ≈ 0. After regularized by our APR, these pixel-level constraints become feasible and stable pixel-wise CORE loss functions that recover polar estimation from the collapse (the blue histogram).

Our main contributions are as follows:

- We devise CORE loss functions from the pixel-level constraints of depth-normal orthogonality. Without any pre-training or surface normal candidates, CORE loss functions enable pixel-wise surface normal regression in a weakly supervised manner by only ground-truth depth map.
- We propose SANE to collaborate with CORE loss functions for multi-task learning. SANE can be easily plugged to existing encoder-decoder approaches of depth estimation without breaking their integrity, which meanwhile benefits the whole performance.
- We achieve steady depth enhancement and superior surface normal prediction. For depth estimation, the experimental results show new state-of-the-art performance across all metrics on NYUv2 and ScanNetv2, and comparable performance on KITTI. For surface normal estimation, despite of weak supervision, it is even comparable with relevant supervised methods.

2. Related Work

Monocular depth estimation. Monocular depth estimation describes a problem that takes a single RGB image as

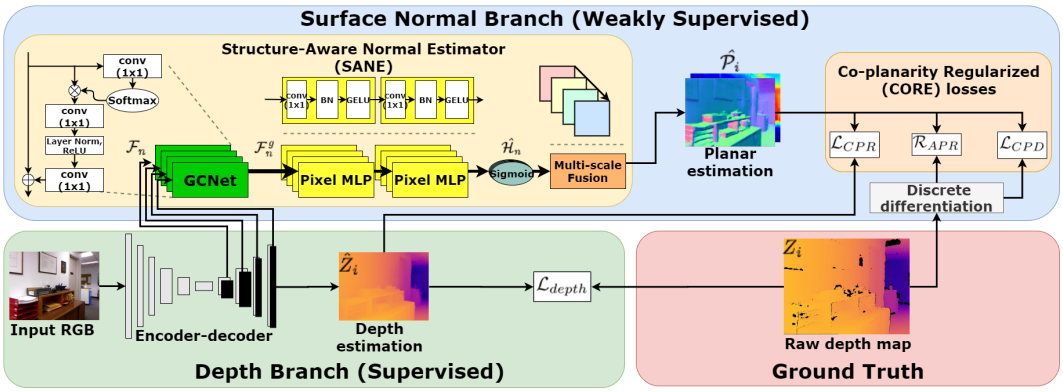


Figure 3: The architecture overview. $\hat{\mathcal{P}}_i$, $\hat{Z}_i$ and Z_i are planar prediction, depth prediction and raw depth map per pixel i, respectively. The planar prediction contains surface normal and distance prediction. The index of multi-scales is donated by n. $\mathcal{F}_n$ are decoder features. $\mathcal{F}_n^g$ are decoder features after the global self-attention. $\hat{\mathcal{H}}_n$ is the intermediate planar prediction at spherical polar coordinates. Discrete differentiation is usually achieved by Sobel operator. $\mathcal{L}_{depth}$ and ($\mathcal{R}_{APR}$, $\mathcal{L}_{CPD}$, $\mathcal{L}_{CPR}$) are depth loss and components of CORE losses explained in Sec. 3.2. The details for SANE are described in Sec. 3.3.

input and predicts the corresponding depth map as output. Recently monocular depth estimation has been extensively studied by DNN approaches [12, 11, 27, 23, 15, 21, 2]. Most of them thoroughly or partially formulate the solutions as an encoder-decoder architecture similar to [36, 13]. Some approaches [6, 14] treat monocular depth estimation as a classification task rather than a regression task. Other methods [24, 19] apply the co-planarity constraint to up-sampling layers or attention layers. As the rising of Transformer architectures [10, 28], the more recent approaches [5, 35, 43, 26, 1] further improve the depth estimation performance by better global perception.

Monocular surface normal estimation. Monocular surface normal estimation defines a similar problem as the above monocular depth estimation. Early attempts mostly follow the similar DNN regression as monocular depth estimation. Recent approaches [18, 9, 39] try to employ extra priors to enhance previous approaches. Specifically, Huang *et al*. [18] took the canonical frames and principal directions as additional supervised signals. Do *et al*. [9] designed a spatial rectifier to learn from the the gravity-aligned training data. VPLNet [39] and self-supervised StructDepth [25] refer to Manhattan world assumptions, *e.g*., Manhattan lines. Besides extra priors, Bae *et al*. [3] proposed a loss function under angular vonMF distribution to predict surface normal and its uncertainty.

Monocular geometry estimation. Monocular geometry estimation combines the monocular depth and surface normal estimation together as a whole 3D geometry, and considers the consistency between them for mutual enhancement [38, 32, 33, 34, 41, 29, 31, 4]. Most methods require both ground truth of depth and surface normals [38, 32, 33, 34, 29, 4]. Qi *et al*. in their works [32, 33] proposed an iterative CNN framework to refine depth map and surface normals from coarse initials. The similar ideas about pre-training and refining are also proposed by SURGE [38], SharpNet [34] and IronDepth [4]. Due to the difficulties on acquisition of the ground truth, there are also works that depend on less or none of ground-truth normals [41, 31]. Among them, Yin *et al*. [41] introduced global sampling for virtual surface normals. Long *et al*. [29] proposed better 3D region sampling and area adaption. The latest work by Patil *et al*. [31] introduces offset vector field and mean plane loss that employs least square fitting. These methods are mostly related to our method, whereas our method is distinguished by the pixel-wisely end-to-end manner of weak supervision.

3. Method

In this section, we explain the details of our methodology. Firstly, we introduce the mathematical model under spherical coordinates, which formulates the pixel-level depth-normal constraints derived from local co-planarity assumption and pinhole camera model. From these constraints, we analytically demonstrate degeneracies and the causes from the view of loss functions. Secondly, APR term is proposed as a countermeasure to the degeneracies, and CORE loss functions are introduced accordingly. At last, we explain details of SANE that collaborates with our loss functions for multi-task learning.

3.1. Spherical Depth-Normal Model

Depth-normal constraint at each individual pixel. We assume the local linearity for common 3D scenes, and it depicts a small tangent plane and its corresponding unit surface normal $\mathbf{n} = (n_1, n_2, n_3)$ with $\|\mathbf{n}\|_2 = 1$ at a 3D point $\mathbf{P_0} = (X_0, Y_0, Z_0)$. (u, v) is the pixel coordinate that $\mathbf{P_0}$ is projected to by pinhole camera model. Given that a surface normal $\mathbf{n}$ is a unit vector, it is a common practice to describe this vector $\in \mathbb{R}^3$ in concise spherical coordinates

$(\theta, \phi) \in \mathbb{R}^2$. We rewrite the surface normal at spherical polar coordinates $\mathbf{n} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$, where $\theta \in [-\pi/2, \pi/2]$ and $\phi \in (-\pi, \pi]$ mean the polar and azimuthal angle, respectively. Moreover, we treat the world coordinates as the same as the camera coordinates, and consider the camera model as an ideal pinhole model in the following sections. The constraint between depth map and surface normals at each individual pixel is obtained by

$$\frac{\sin\theta\cos\phi}{n_4(\theta,\phi)}\frac{u-c_x}{f_x} + \frac{\sin\theta\sin\phi}{n_4(\theta,\phi)}\frac{v-c_y}{f_y} + \frac{\cos\theta}{n_4(\theta,\phi)} = Z^{-1}, \tag{1}$$

where Z^{-1} is the inverse depth map, and $n_4 = \sin\theta\cos\phi X_0 + \sin\theta\sin\phi Y_0 + \cos\theta Z_0$ is the distance from world origin to the tangent plane. Conditionally on Z and (u, v), n_4 can be parameterized by θ and ϕ, namely $n_4(\theta, \phi)$. (c_x, c_y) and (f_x, f_y) are the camera principal points and focal lengths, respectively. The fundamental Eq. (1) associates surface normals and depth map at each individual pixel. Minimization on pixel-wise difference between the left-hand side (LHS) and the right-hand side (RHS) of Eq. (1) suggests a loss objective, *e.g.*, pixel-wise L1 or L2 distance.

The trivial solution of Eq. (1). However, without considering neighboring pixels, Eq. (1) indicates a quick trivial solution ($\theta \equiv 0, n_4 \equiv Z$). This trivial solution degenerates the estimation of surface normals constrained by Eq. (1), since the back-propagation with gradient-decent strategy stably converges the prediction to it as a shortcut. We notice that $\theta \equiv 0$ is the persistent suboptimal state that incurs polar angle collapse and meanwhile causes the trivial solution.

Depth-normal constraint between local pixels. The first-order partial derivatives at pixel level connect an individual pixel with local neighbors, which prevents the emergence of the trivial solution of Eq. (1). Naturally, we take the first-order partial derivatives of u and v on both sides of Eq. (1). The depth-normal constraint between local pixels is formulated by

$$\left(\frac{\sin\theta\cos\phi}{n_4(\theta,\phi)f_x}, \frac{\sin\theta\sin\phi}{n_4(\theta,\phi)f_y}\right) = \nabla(Z^{-1}), \tag{2}$$

where $\nabla(Z^{-1})$ is the 2D image gradient of inverse depth map. Even though Eq. (2) does not suffer from the trivial solution anymore, common objectives, *e.g.*, L1 or L2, are still not applicable for Eq. (2). The reason lies in the imbalanced distribution of un-normalized $\nabla(Z^{-1})$. Intuitively, for the loss objective, cosine similarity would be effective with only angular differences taken into consideration.

The degeneracy of Eq. (2). However, cosine similarity introduces another degeneracy on Eq. (2). Given that f_x and f_y are approximate constants and independent term of $\sin\theta$, we simplify Eq. (2) to the following new relation:

$$\sin\theta\left(\frac{\cos\phi}{n_4(\theta,\phi)}, \frac{\sin\phi}{n_4(\theta,\phi)}\right) \sim \nabla(Z^{-1})$$

From the above relation, it is not difficult to realize that $\sin\theta$, as a proportion, hardly contributes to minimize cosine similarity. The consequence is that optimization of θ quickly becomes inactive around the suboptimal state $\theta \approx 0$. The pixel-level constraint from Eq. (2) loses the degree of freedom of θ, which also causes collapse on polar angle estimates. Therefore, the objective of cosine similarity degenerates the optimization of (θ, ϕ, n_4) as well.

3.2. Co-planarity Regularized (CORE) Losses.

Hereafter we specify the arbitrary pixel (u, v) as i for the brief and consistent notation. Planar estimation $\hat{\mathcal{P}}_i$ is $(\hat{n}_1, \hat{n}_2, \hat{n}_3, \hat{n}_4)$ [1] parameterized by spherical polar predictions: $(\hat{\theta}, \hat{\phi})$. We use the notation of planar estimation in our CORE losses for brevity.

Adaptive polar regularization (APR). According to the above analysis on the degeneracies, the fundamental cause is the polar angle collapse during back-propagation. The intuitive but crucial countermeasure becomes how to regularize the predicted $\hat{\theta}$ to be away from its suboptimal state $\hat{\theta} \approx 0$. To this end, we propose APR that consists of a polar regularizer g_i and an adaptive weight map ω_i. The polar regularizer g_i exploits penalties to prevent the polar angle collapse. It is simply parameterized by only $\hat{\theta}$, which is defined as:

$$g_i = -\ln(4\hat{n}_3^{\frac{1}{4}}(1 - \hat{n}_3^{\frac{1}{4}})), \tag{3}$$

where scalar 4 aims to arrange penalties $\in [0, \infty)$, and the power $1/4$ on $\hat{n}_3$[1] is to set regularization minimums to close to $\pm\pi/2$ (see more discussion on the choice of power in our supplementary material). As a result, regularized depth-normal constraints tend to guide polar angle estimation along the decaying direction of regularization in range of $(0^-, -\pi/2)$ and $(0^+, \pi/2)$ which are also consistent to the real value domain of $\hat{\theta}$. Polar regularizer g_i forces $\hat{\theta}$ away from 0, but $\hat{\theta} \approx 0$ shall reasonably exist in the 3D scenes as planes parallel to the image XY-plane. Accordingly, as the other component of APR, we introduce adaptive weight map ω_i to control the application of polar regularization per pixel:

$$\omega_i = 1 - \exp\left(-\frac{\|\nabla(Z^{-1})_i\|_2^2}{\gamma\sigma^2}\right), \tag{4}$$

where σ is $std(\|\nabla(Z^{-1})\|_2)$ for all the pixels from a mini-batch, and γ is a scalar hyper-parameter for different

[1] $\hat{n}_1 = \sin\hat{\theta}\cos\hat{\phi}$, $\hat{n}_2 = \sin\hat{\theta}\sin\hat{\phi}$, $\hat{n}_3 = \cos\hat{\theta}$

datatets. The purpose of ω_i is to utilize the statistical knowledge of $\|\nabla(Z^{-1})_i\|_2 \in \mathbb{R}$ by a Gaussian RBF kernel, which adaptively loosens the polar regularizer and maintains the local correlation. Finally, we formulate APR as:

$$\mathcal{R}_{APR} = \frac{1}{T}\sum_i(\omega_i \odot g_i), \tag{5}$$

where $\odot$ is the element-wise multiplication, and T is hereafter the total number of pixels having valid raw depth values ($Z_i \neq 0$). $\nabla(Z^{-1})_i$ can be approximated by a 3×3 Sobel operator on the inverse raw depth. We find that $\gamma = 4$ and $\gamma = 16$ are empirically working well for indoor and outdoor datasets that we use in Sec. 4, respectively.

Co-planar differentiation loss (CPD loss). With the above APR that resolves the optimization degeneracies, we firstly introduce our designed CPD loss function according to Eq. (2), which minimizes the angular difference between $\hat{\mathbf{q}}_i = (\hat{n}_1/(\hat{n}_4 f_x), \hat{n}_2/(\hat{n}_4 f_y))$[1] and $\mathbf{q}_i = \nabla(Z^{-1})_i$. To be sensitive to small angular errors, we tailor the cosine similarity to the angular loss. Additionally, inspired by Do *et al*. [9] who suggest that a normalized L2 loss at large angular error ($> \pi/2$) is more robust to outliers, we segment our CPD loss function into two parts as follows:

$$\mathcal{L}_{CPD} = \begin{cases} \dfrac{1}{N}\sum_i(s_i^{-1}) & s_i \in [0, 1], \\ \dfrac{1}{M}\sum_i(\pi/2 - s_i) & s_i \in [-1, 0), \end{cases} \tag{6}$$

where s_i is the cosine similarity between $\hat{\mathbf{q}}_i$ and $\mathbf{q}_i$, and s_i^{-1} is the inverse cosine. $\hat{\mathbf{q}}_i$ is related to network predictions and focal lengths. N and M are the total number of valid pixels that satisfy $s_i \in [0, 1]$ and $s_i \in [-1, 0)$, respectively, where $N + M = T$.

Co-planar refinement loss (CPR loss). Next, we propose our CPR loss which aims to compensate normal estimation bias and meanwhile refine depth map and surface normals together. Although CPD and APR can guide surface normals regression by the first-order depth-normal constraint, they discard the constant term per pixel between absolute and relative estimates. To compensate these, we consider the absolute depth-normal constraint from Eq. (1) for refinement. As a result, according to Eq. (1), we design CPR loss in terms of the smooth L1 that is effective in refinement tasks.

$$\mathcal{L}_{CPR} = \frac{1}{T}\sum_i |\mathbf{c_i} \cdot \hat{\mathbf{p}}_\mathbf{i} - \hat{Z}_i^{-1}|_{smooth}, \tag{7}$$

where $\mathbf{c_i} = ((u - c_x)/f_x, (v - c_y)/f_y, 1)$ is calculated from image coordinates and camera intrinsics. $\hat{\mathbf{p}}_\mathbf{i} = (\hat{n}_1/\hat{n}_4, \hat{n}_2/\hat{n}_4, \hat{n}_3/\hat{n}_4)$[1] and $\hat{Z}_i^{-1}$ are both from network predictions.

Overall loss. Finally, our overall loss function is given as:

$$\mathcal{L}_{total} = \underbrace{(\mathcal{L}_{CPD} + \mathcal{R}_{APR} + \mathcal{L}_{CPR})}_{\mathcal{L}_{CORE}} + \mathcal{L}_{depth}, \tag{8}$$

where $\mathcal{L}_{depth}$ is a loss to supervise the depth estimation. Here we use Scale-Invariant Logarithmic (SILog) loss [12].

3.3. Structure-Aware Normal Estimator (SANE)

The overall architecture shown in Fig. 3 contains two major components, namely an encoder-decoder architecture for depth estimation and our designed SANE for surface normal estimation. The design for SANE as shown in upper part of Fig. 3 contains three cascade blocks to perform global perception, pixel-wise regression and multi-scale fusion successively.

Global perception. We propose to utilize Global Context Networks (GCNet) [7] to enhance global perception, since the multi-scale features $\mathcal{F}_n$ could have insufficient global information. GCNet introduces a light-weighted way to pixel-wisely apply global self-attention at feature level. It can be conveniently inserted after feature maps that are extracted from the encoder-decoder architecture.

Pixel-wise regression. We design simple pixel-wise Multi-Layer Perceptions (MLP) for the regression task. The Pixel MLP is easily achieved by pixel-wise Conv1x1 + BN + GELU for channels. The regression contains 2 Pixel MLPs that squeeze the channels of $\mathcal{F}_n^g$: $C_n \times H_n \times W_n$ by the ratio of $\lambda = 0.5$, and then outputs the intermediate prediction $\mathcal{H}_n$: $3 \times H_n \times W_n$ that contains $\hat{\theta}, \hat{\phi}, \hat{n}_4$ as channel components. We apply Sigmoid on $\hat{\mathcal{H}}_n$ and rectify it to the actual value domains by scalar $\hat{\theta} \in [-\pi/2, \pi/2]$, $\hat{\phi} \in (-\pi, \pi]$ and $\hat{n}_4 \in [Z_{min}, Z_{max}]$.

Multi-scale fusion. We fuse the planar prediction $\hat{\mathcal{P}}_{n,i}$ that contains $(\hat{n}_1, \hat{n}_2, \hat{n}_3)$[1] and $\hat{n}_4$ at the pixel i per scale n. For the fusion strategy, we bi-linearly interpolate $\hat{\mathcal{P}}_{n,i}$ to the original size, and pixel-wisely assemble them over scales. Specifically, for $(\hat{n}_1, \hat{n}_2, \hat{n}_3)$[1] $\in \mathbb{R}_3$, we pixel-wisely apply weighted summation over scales, *e.g.*, from $1/4$ scale to $1/32$ scale, and normalization after each summation. The weights are related to the up-ratio as $1/upratio$. As for $\hat{n}_4$, it is the common pixel-wisely average over all scales.

4. Experiments

4.1. Implementation Details

In this work, we choose existing depth estimation methods, namely BTS [24], Adabins [5] and NeWCRs [43], to present the effectiveness of CORE losses and SANE on various DNN architectures. For details, BTS proposes a

Method	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	Abs.Rel↓	RMSE↓	$\log_{10} \downarrow$
Eigen *et al.* [12]	0.769	0.950	0.988	0.158	0.641	–
DORN [14]	0.828	0.965	0.992	0.115	0.509	0.051
GeoNet [32]	0.862	0.965	0.989	0.113	0.527	0.049
VNL [41]	0.875	0.976	0.994	0.108	0.416	0.048
SharpNet* [34]	0.888	0.979	0.995	0.139	0.495	0.047
ASNDepth [29]	0.890	0.982	0.996	0.101	0.377	0.044
DPT* [35]	0.904	0.988	**0.998**	0.110	0.357	0.045
P3Depth [31]	0.904	0.988	**0.998**	0.104	0.356	0.043
IronDepth [4]	0.910	0.985	0.997	0.101	0.352	0.043
BinsFormer [26]	0.925	0.989	0.997	0.094	0.330	0.040
PixelFormer [1]	0.929	0.991	**0.998**	0.090	0.322	0.039
BTS [24]	0.885	0.978	0.994	0.110	0.392	0.047
Adabins [5]	0.903	0.984	0.997	0.103	0.364	0.044
NeWCRFs [43]	0.922	**0.992**	**0.998**	0.095	0.334	0.041
Ours (BTS)	0.890	0.982	0.996	0.106	0.375	0.046
Ours (Adabins)	0.899	0.984	0.997	0.106	0.359	0.044
Ours (NeWCRFs)	**0.932**	**0.992**	**0.998**	**0.088**	**0.317**	**0.038**

Table 1: Quantitative results of depth estimation on NYUv2. "*" means using additional data during training.

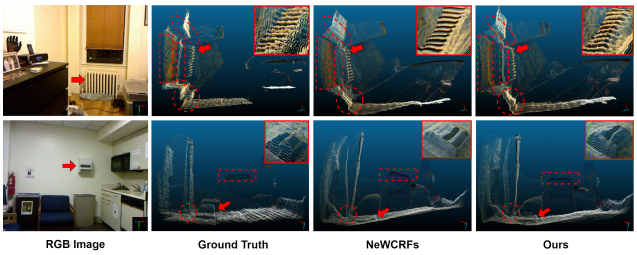


Figure 4: The point cloud from bird view. The dashed box indicates the better flatness; The dashed circle indicates the better plane intersection; The arrow and zoom-in box indicate better recovered surface. Zoom in for better view.

CNN-based encoder-decoder architecture. Adabins combines a CNN-based encoder-decoder with a Vision Transformer [10]. The latest NeWCRFs employs the Swin-Transformer [28] in the design. For fair comparison, except our CORE losses and SANE, other implementation details, such as learning rate, batch size, training epoches and etc., are identical to the original methods. Relevant details are omitted because of the limited space.

4.2. Datasets

NYU depth v2 [37] is an indoor RGB-D dataset which provides samples with the resolution of 480×640 and maximum depth of 10 meters. We train our network with the same subsets as [24, 5, 43], and evaluate our method on the official 654 testing samples with center cropping defined by Eigen *et al.* [12]. For surface normal evaluation, we use 654 ground truth generated by Ladicky *et al.* [22] on only the valid pixels by following Zhang *et al.* [44] and Do *et al.* [9].

KITTI [16] is a stereo-LiDAR dataset for outdoor scenes. It provides the 376×1241 stereo-RGBs and sparse depth map with maximum depth of 80 meters. We reserve the training subset defined by [24, 43]. The test set contains

Method	Sup.	Mean↓	11.2° ↑	22.5° ↑	30° ↑
Surface Normal Estimation Network					
Ladicky *et al.* [22]	✓	35.5	24.0	45.6	55.9
Wang *et al.* [40]	✓	28.8	35.2	57.1	65.5
Eigen *et al.* [11]	✓	23.7	39.2	62.0	71.1
3D Geometry Estimation Network					
GeoNet [32]	✓	36.8	15.0	34.5	46.7
VNL [41]	×	24.6	34.1	60.7	71.7
ASNDepth [29]	✓	**20.0**	43.5	69.1	**78.6**
IronDepth [4]	✓	20.8	**49.7**	**70.5**	77.9
Calculated Surface Normal from Depth					
DORN [14]	×	36.6	15.7	36.5	49.9
Hu *et al.* [17]	×	32.1	24.7	48.5	59.9
BTS [24]	×	44.0	14.4	32.2	43.2
Adabins [5]	×	33.2	22.3	47.2	58.7
NeWCRFs [43]	×	29.5	27.1	52.2	64.5
Ours (BTS)	×	30.2	24.6	47.4	58.7
Ours (Adabins)	×	29.0	25.5	48.1	60.1
Ours (NeWCRFs)	×	*21.9*	*34.6*	*63.6*	*75.4*

Table 2: Quantitative results of surface normal estimation on NYUv2. "✓" means training (including pre-training) supervised by ground-truth normals, while "×" means not.

697 samples specified by Eigen *et al.* [12] and cropped as Garg *et al.* [15].

ScanNetv2 [8] is also an RGB-D video dataset containing 2.5 million views in more than $1,500$ scans indoors. We use the $2,167$ official test samples to cross-evaluate our method that is well trained with NYUv2 dataset. This dataset is not used as training data.

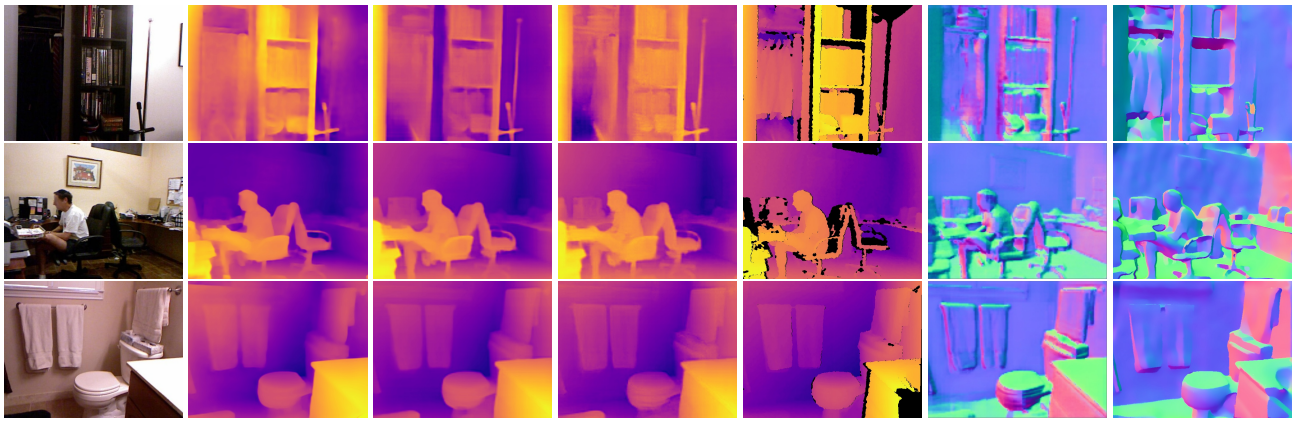
4.3. Evaluation Metrics

Depth. We consider evaluation metrics used in Eigen *et al.* [12] to compare our depth with others. These metrics are defined as: Root Mean Squared Error (RMSE) and its *log* variant RMSE_{log}, Average Relative Error (Abs.ReL) and its square variant Sqr.Rel, Average ($\log_{10}$) Error, and Threshold Accuracy (δ_i) for 1.25, 1.25^2, 1.25^3.

Surface normal. Following previous works [25, 42], the main ranking metrics of mean and the percentage of pixels with error under thresholds $t \in [11.25°, 22.5°, 30°]$ are reported for angular errors.

4.4. Comparison with the State-of-the-Art

Evaluation on NYU depth v2. We evaluate our method on NYUv2 dataset for indoor scenes. As presented in Tab. 1, the performance of depth estimation has been steadily improved after applying CORE losses and SANE onto the original methods [24, 5, 43]. Since NYUv2 dataset has been saturated for a while [43], we use relative gains for better indication. On NeWCRFs [43], our method further improves its main indicators by 5.1% (RMSE) and 7.3% (AbsRel). Also, the enhanced NeWCRFs [43] by our method outperforms the latest state-of-the-art of PixelFormer [1] on all the metrics. Besides the metrics, our method is able to ameliorate the 3D geometry estimation from over-smoothing,



(a) Input RGB. (b) Adabins. (c) NeWCRFs. (d) Our depth. (e) GT depth. (f) Our normals. (g) GT normals.

Figure 5: The qualitative results on NYUv2. Depth and normal estimation are from ours (NeWCRFs) because of limited spaces. More qualitative results can be found in our supplementary material. Zoom-in and best view in color.

Method	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	Abs.Rel↓	RMSE↓	Sqr.Rel↓	$\text{RMSE}_{log} \downarrow$
Eigen *et al.* [12]	0.702	0.898	0.967	0.203	6.307	1.548	0.282
DORN [14]	0.932	0.984	0.994	0.072	2.727	0.307	0.120
GeoNet [33]	0.897	0.968	0.986	0.094	–	–	–
VNL [41]	0.938	0.990	0.998	0.072	3.258	–	0.117
DPT* [35]	0.959	0.995	**0.999**	0.062	2.573	–	0.092
P3Depth [31]	0.953	0.993	0.998	0.071	2.842	0.270	0.103
BinsFormer [26]	0.974	**0.997**	**0.999**	0.052	2.098	0.151	0.079
PixelFormer [1]	**0.976**	**0.997**	**0.999**	**0.051**	**2.081**	**0.149**	**0.077**
BTS [24]	0.956	0.993	0.998	0.059	2.756	0.245	0.096
Adabins† [5]	0.962	0.995	**0.999**	0.058	2.422	0.217	0.091
NeWCRFs [43]	0.974	**0.997**	**0.999**	0.052	2.129	0.155	0.079
Ours (BTS)	0.962	0.994	**0.999**	0.060	2.442	0.202	0.092
Ours (Adabins)	0.962	0.995	**0.999**	0.059	2.379	0.195	0.091
Ours (NeWCRFs)	**0.976**	**0.997**	**0.999**	0.052	2.095	**0.149**	**0.077**

Table 3: Quantitative results of depth estimation on KITTI. "†" denotes the retrained model with the same training set as [24, 43]. "*" means using additional training data.

which is visualized by point clouds in Fig. 4. These point clouds present that our method can rectify curved planes, sharpen plane intersection, and recover local surface with more precise shape and texture, *e.g.*, the hanging box and heater. Our depth estimation is visualized in Fig. 5, which shows similarly better geometry details.

For surface normal estimation, Tab. 2 demonstrates that our weakly supervised method outperforms relevant methods that involve none of ground-truth normals. The comparison results are mainly cited from the original papers and published articles [41, 29]. For NeWCRFs [43] and Adabins [5], we compute the surface normals from their estimated depth by least-square plane fitting, following the common practice. Although ASNDepth [29] and IronDepth [4] provide slightly better performance, they explicitly require the ground-truth normals. From Fig. 5, qualitative results show that our method is able to predict more geometry details of textures and boundaries, which are even absent in the ground truth, *e.g.*, the thin pole (in the first row) and wrinkles on the towels (in the third row).

Method	$\sigma_1 \uparrow$	$\sigma_2 \uparrow$	$\sigma_3 \uparrow$	Abs.Rel↓	RMSE↓	$\log_{10} \downarrow$
VNL [41]	0.565	0.856	0.957	0.238	0.505	0.105
ASNDepth [29]	0.609	0.861	0.955	0.233	0.484	0.100
P3Depth [31]	0.551	–	–	0.223	0.538	–
BTS [24]	0.583	0.858	0.951	0.246	0.506	0.104
Adabins [5]	0.622	0.858	0.945	0.223	0.466	0.100
NeWCRFs [43]	0.727	0.943	0.984	0.182	0.359	0.081
Ours (BTS)	0.629	0.888	0.966	0.224	0.450	0.094
Ours (Adabins)	0.642	0.885	0.963	0.218	0.443	0.093
Ours (NeWCRFs)	**0.739**	**0.946**	**0.986**	**0.178**	**0.349**	**0.073**

Table 4: Quantitative results of depth estimation on the ScanNetv2 (cross-evaluation on 2,167 official test images).

Evaluation on KITTI. We only evaluate our depth estimation on KITTI dataset for outdoor scenes because of the absence of standard ground-truth normals. Tab. 3 shows that our method still progressively enhances the original depth estimation. On BTS [24], our method considerably reduces its RMSE by 11.4% and Sqr.Rel by 17.6%. On NeWCRFs [43], our method boosts its performance very close to the latest PixelFormer [1]. It is worth noting that this performance gain requires no extra network parameters, because SANE can be discarded after training. Moreover, on the condition of distance, we observe that our method consistently ameliorates Abs.Rel for near scenes < 20m, and yet slightly affects Abs.Rel at far scenes > 50m (see details in our supplementary material). We argue that the local co-planarity assumption holds mostly for near regular scenes, while some far scenes are exceptions. Fortunately, the depth density in far range > 50m is usually low, and the overall impact is insignificant.

Evaluation on ScanNetv2. To indicate the generalisation of our method, we perform a cross-evaluation on ScanNetv2 dataset, similar to [29, 31]. Tab. 4 shows that our method steadily improves corresponding original approaches without decreasing the generalisation. Also, our NeWCRFs-based solution states the best results.

Depth loss	CORE losses			Depth	Normal
	APR	CPD loss	CPR loss	RMSE↓	Mean↓
	✓	✓	✓	–	35.4
✓			✓	0.331	✗
✓		✓		0.324	✗
✓		✓	✓	0.324	✗
✓	✓	✓		0.319	27.6
✓	✓	✓	✓	**0.317**	**21.9**

Table 5: Ablation studies on our proposed CORE losses. ✗ means a failure prediction because of the polar angle collapse. "-" means not applicable. Without depth loss, CPR loss uses ground-truth depth instead.

Method	Backbone	GCNet	Depth	Normal
			RMSE↓	Mean↓
BTS [24]	DenseNet	×	0.380	34.9
		✓	0.375	30.2
Adabins [5]	EfficientNet	×	0.363	32.1
		✓	0.359	29.0
NeWCRFs [43]	Swin-T	×	0.317	22.6
		✓	0.317	21.9

Table 6: Ablation studies on our proposed SANE. The global perception is crucial to benefit the surface normal estimation for CNN architectures.

4.5. Ablation Study

In this section, we conduct ablation studies to analyze the effectiveness of our design. The ablation studies are mainly performed with NeWCRFs [43] and NYUv2 [37] as default.

CORE losses. 1) APR is the spotlight of our design for CORE losses, which guarantees the weakly supervised normal estimation. As discussed in Sec. 3.1 and shown in Fig. 2, without APR, surface normals could not be properly predicted because of the polar angle collapse. 2) According to Tab. 5, we notice that CPD loss still improves the performance of depth prediction regardless of APR. The latent reason is that azimuthal of ϕ is persistently optimized by CPD loss even under the degeneracy. 3) The results in Tab. 5 also indicate that CPR loss is crucial. This loss further polishes both performance, especially for the surface normals. As discussed in Sec. 3.2, CPR loss compensates and refines the relative surface normal estimation, so that the estimates are more consistent to the ground truth as shown in Fig. 6. In a nutshell, each component of our CORE losses is indispensable (see more analysis in our supplementary material).

SANE. As presented in Tab. 6, when GCNet [7] is removed from SANE, surface normal estimation obviously decreases on performance for BTS [24] and Adabins [5] which employ CNN backbones, whereas the impact is slight for NeWCRFs [43]. The reason lies in that NeWCRFs has already utilized Swin-Transformer [28] to perform decent global perception. This fact suggests that SANE can benefit the surface normal estimation for some CNN architectures without sufficient global information. Moreover, the results in Tab. 7 indicate that Pixel MLPs and Multi-scale Fusion basically remains stable to the configurations except for the extreme setting, *e.g.* 1 Pixel MLP and 1 Multi-scale Fusion. Our default settings aim to balance the accuracy and computation.

(a) Input RGB. (b) w/o CPR. (c) with CPR. (d) GT.

Figure 6: The effectiveness of CPR loss on surface normal estimation. The surface normal estimates are rectified and polished by CPR loss. Zoom-in and best view in color.

Module	Configs	Depth	Normal
		RMSE↓	Mean↓
Pixel MLPs	1	0.334	34.0
	2	**0.317**	**21.9**
	3	0.317	21.8
Multi-scale Fusion	1	0.319	24.2
	2	0.317	22.7
	4	**0.317**	**21.9**

Table 7: More ablation studies on our proposed SANE. The default settings are in bold font.

Multi-task learning. As also presented in Tab. 5, multi-task learning is important for both depth and surface normal estimation. Particularly, the surface normal estimation is boosted by 13.5° after depth branch is enabled. Thus, it is mostly a wise choice to perform such full multi-task training and then reserve the necessary regression head at inference time, because multi-task learning with SANE is light-weighted and involved only in the training stage.

5. Conclusion

In this paper, we explore the feasibility to utilize depth-normal constraints as pixel-wise loss functions for 3D geometry estimation. Firstly, we propose CORE losses from spherical depth-normal constraints, which enable surface normal regression weakly supervised by only ground-truth depth map. Then, we design SANE for multi-task learning, which introduces CORE losses to existing methods of depth estimation, and boosts the whole performance. Finally, our method achieves new state-of-the-art of indoor depth estimation, comparable outdoor depth estimation and superior surface normal estimation. In future, we will focus on the further improvement of our method for outdoor scenes.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5861–5870, 2023. 3, 6, 7

[2] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. In *arXiv preprint arXiv:1812.11941*, 2018. 3

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2022. 1, 2, 3, 6, 7

[5] 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 (CVPR)*, 2021. 3, 5, 6, 7, 8

[6] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. 3

[7] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 5, 8

[8] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 6

[9] Tien Do, Khiem Vuong, Stergios I. Roumeliotis, and Hyun Soo Park. Surface normal estimation of tilted images via spatial rectifier. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 5, 6

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 3, 6

[11] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 3, 6

[12] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2014. 3, 5, 6, 7

[13] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Phili Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2758–2766. 3

[14] Huan Fu, Mingming Gong, ChaohuiWang, Nematollah Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018. 3, 6, 7

[15] Ravi Garg, Vijay Kumar B.G., Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–756, 2016. 3, 6

[16] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 32:1231–1237, 2012. 1, 6

[17] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051, 2018. 2, 6

[18] Jingwei Huang, Yichao Zhou, Thomas Funkhouser, and Leonidas Guibas. Framenet: Learning local canonical frames of 3d surfaces from a single rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[19] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkila. Guiding monocular depth estimation using depth-attention volume. In *arXiv preprint arXiv:2004.02760*, 2020. 3

[20] Uday Kusupati, Shuo Cheng, Rui Chen, and Hao Su. Normal assisted stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[21] Yevhen Kuznietsov, Jörg Stückler, and Bastian Leibe. Semi-supervised deep learning for monocular depth map prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2215–2223, 2017. 3

[22] Lubor Ladicky, Bernhard Zeisl, and Marc Pollefeys. Discriminatively trained dense surface normal estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1, 2, 6

[23] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2016. 3

[24] 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. In *arXiv preprint arXiv:1907.10326*, 2019. 3, 5, 6, 7, 8

[25] Boying Li, Yuan Huang, Zeyu Liu, Danping Zouy, and Wenxian Yu. Structdepth: Leveraging the structural regu-

larities for self-supervised indoor depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 6

[26] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 3, 6, 7

[27] Fayao Liu, Chunhua Shen, Guosheng Lin, and I. Reid. Learning. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:2024–2039, 2015. 3

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 8

[29] Xiaoxiao Long, Cheng Lin, Lingjie Liu, and Wei Li. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 6, 7

[30] Xiaoxiao Long, Lingjie Liu, Christian Theobalt, and Wenping Wang. Occlusion-aware depth estimation with adaptive normal constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[31] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 7

[32] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 283–291, 2018. 2, 3, 6

[33] Xiaojuan Qi, Zhengzhe Liu, Renjie Liao, Philip HS Torr, Raquel Urtasun, and Jiaya Jia. Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 3, 7

[34] Michaël Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2019. 2, 3, 6

[35] Réne Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 6, 7

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. 3

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 6, 8

[38] Peng Wang, Xiaohui Shen, Bryan Russell, Scott Cohen, Brian Price, and Alan Yuille. Surge: Surface regularized geometry estimation from a single image. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2016. 2, 3

[39] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[40] Xiaolong Wang, David Fouhey, and Abhinav Gupta. Designing deep networks for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[41] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 3, 6, 7

[42] Zehao Yu, Lei Jin, and Shenghua Gao. P^2net: Patch-match and plane-regularization for unsupervised indoor depth estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[43] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neuralwindow fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 6, 7, 8

[44] Y. Zhang, S. Song, E. Yumer, M. Savva, J.Y. Lee, H. Jin, and T Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6