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

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

In this supplementary material, we provide extra explanation and visualization for a better understanding of our paper. Specifically, Appendix A provides more detailed derivations of the proposed spherical depth-normal model. By Appendix B, Appendix C and Appendix D, we then present extra analysis and explanation on each component of our CORE losses, and further indicate their indispensability. In Appendix E, we evaluate our method on KITTI dataset, conditioned on distance. This experiment indicates that our method can make use of the depth-normal constraints to improve depth performance in near scenes, and yet affect depth estimation in far scenes insignificantly. In Appendix F, we present histogram comparison between depth estimates and ground truth depth, which indicates the effectiveness of our method on refining 3D geometry estimation. In Appendix G, additional qualitative results are presented in addition to those shown in our main manuscript. Appendix H shows some failure cases of surface normal prediction in outdoor scenes, which will be further investigated as our future work. As the same as our main manuscript, notations with hat mean corresponding estimates, *e.g.*, $\hat{\theta}$ means the estimation on θ.

A. Derivations of spherical depth-normal model

A.1. Depth-normal association at each individual pixel.

Let's define a point $\mathbf{P_0} = (X_0, Y_0, Z_0)$ on a 3D scene and the corresponding unit surface normal $\mathbf{n} = (n_1, n_2, n_3)$ with $\|\mathbf{n}\|_2 = 1$ of the tangent plane. According to the plane equation, an arbitrary point $\mathbf{P}$ at the tangent plane could be written as:

$$\mathbf{n} \cdot (\mathbf{P} - \mathbf{P_0}) = 0 \tag{1}$$

where $\cdot$ mean the dot product between two vectors. Write Eq. (1) out in components:

$$n_1 X + n_2 Y + n_3 Z = n_4 \tag{2}$$

where $n_4 = n_1 X_0 + n_2 Y_0 + n_3 Z_0$. Recall the pinhole camera model that projects the 3D points $P = (X, Y, Z)$ onto the 2D image plane at pixel coordinates (u, v):

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{X}{Z} \\ \frac{Y}{Z} \\ 1 \end{bmatrix} \tag{3}$$

where, (c_x, c_y) and (f_x, f_y) are camera principal points and focal lengths respectively. Given Z at pixel coordinates as $Z(u, v)$, solving X and Y from Eq. (3) and inserting them into Eq. (2) give:

$$n_1 \frac{u - c_x}{n_4 f_x} + n_2 \frac{v - c_y}{n_4 f_y} + \frac{n_3}{n_4} = Z^{-1}(u, v) \tag{4}$$

where Z^{-1} is the per-pixel reciprocal of the depth map, namely the inverse depth map. We rewrite the unit surface normal $\mathbf{n}$ as spherical polar coordinates:

$$\mathbf{n} = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta), \tag{5}$$

where $\theta \in [-\pi/2, \pi/2]$ and $\phi \in (-\pi, \pi]$ mean the polar and azimuthal angles respectively. Inserting Eq. (5) to Eq. (4) gives:

$$Z^{-1}(u,v) = \frac{\sin\theta\cos\phi}{n_4 f_x}(u - c_x) + \frac{\sin\theta\sin\phi}{n_4 f_y}(v - c_y) + \frac{\cos\theta}{n_4} \tag{6}$$

A.2. Depth-normal association between local pixels.

Eq. (6) could be re-formulated as:

$$\begin{aligned} Z^{-1}(u,v) &= \frac{\sin\theta\cos\phi}{n_4 f_x}(u - c_x) + \frac{\sin\theta\sin\phi}{n_4 f_y}(v - c_y) + \frac{\cos\theta}{n_4} \\ &= \frac{\sin\theta\cos\phi}{n_4 f_x}u + \frac{\sin\theta\sin\phi}{n_4 f_y}v + \underbrace{(\frac{\cos\theta}{n_4} - \frac{\sin\theta\cos\phi}{n_4 f_x}c_x - \frac{\sin\theta\sin\phi}{n_4 f_y}c_y)}_{\text{This term is not related to } u \text{ and } v.} \end{aligned} \tag{7}$$

By applying the partial derivatives of u and v to the both sides of Eq. (7), we have the following equation:

$$(\frac{\sin\theta\cos\phi}{n_4 f_x}, \frac{\sin\theta\sin\phi}{n_4 f_y}) = \nabla(Z^{-1})(u,v) \tag{8}$$

At last, Eq. (6) and Eq. (8) are equation (1) and equation (2) in our main manuscript respectively.

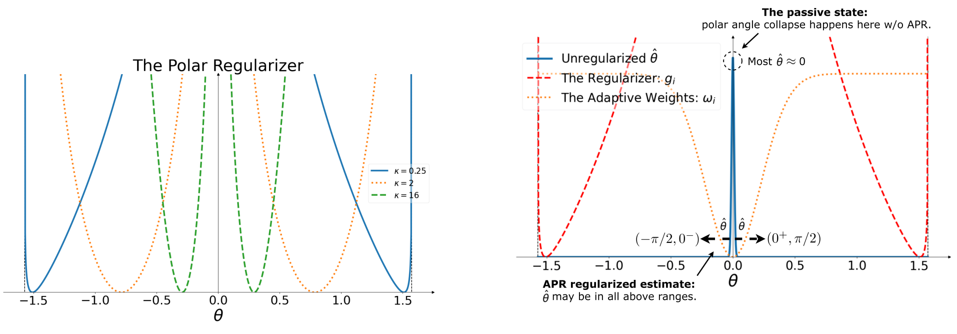
B. More Analysis on APR

B.1. The choice of power for polar regularization

We use power of $1/4$ on $\hat{n}_3$[1] for our polar regularization in our main manuscript, the purpose of which is to push the minimum regularization to around $\pm\pi/2$ for its effectiveness along all the definition domain of $\hat{\theta} \in (-\pi/2, +\pi/2)$. In addition, regularization minimums around $\pm\pi/2$ benefit CPR loss to effectively suppress bias induced by APR (see details in Appendix D.2). To visualize the effect of power on $\hat{n}_3$, let's define the power on $\hat{n}_3$ as κ. The polar regularizer becomes:

$$g_i = -\ln(4\hat{n}_3^{\kappa}(1 - \hat{n}_3^{\kappa})). \tag{9}$$

As presented in Fig. 1a, with decreasing of κ, the minimum regularization is progressively pushed to $\pm\pi/2$. Empirically, $\kappa = 1/4$ is enough to set regularization minimum close to $\pm\pi/2$.



(a) The choice of κ.

(b) The APR.

Figure 1: The choice of power for polar regularization.

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

C. More Analysis on CPD loss

C.1. The distribution of $\nabla(Z^{-1})$

From below Fig. 2, we can see the fact that $\nabla(Z^{-1}) \in \mathbb{R}^2$ has imbalanced data distribution, and the density is extremely high around 0. This data distribution is usually unfriendly to the objectives of L1 or L2. As a result, we employ the cosine similarity (the angular difference) as the objective in our CPD loss.

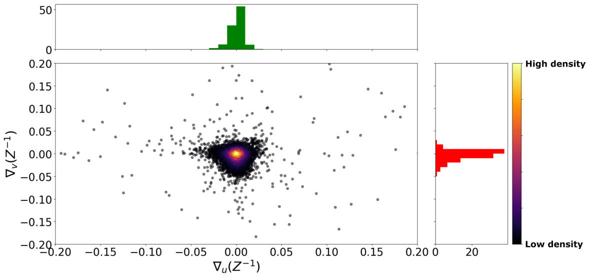


Figure 2: The distribution of $\nabla(Z^{-1})$. This figure is plotted with the valid data ($Z_i \neq 0$) from 16 random depth maps of NYUv2 [6]. The top green histogram is the density of $\nabla_u(Z^{-1})$ and the right red histogram is the density of $\nabla_v(Z^{-1})$.

C.2. Persistent constraint on ϕ by CPD loss

Let's define $\nabla(Z^{-1}) = (d_u, d_v)$. According to Eq. (8) and the definition of CPD loss in our main manuscript, the cosine similarity s_i is calculated as:

$$
\begin{aligned}
s_i &= \frac{\frac{\sin\hat{\theta}\cos\hat{\phi}}{\hat{n}_4 f_x} d_u + \frac{\sin\hat{\theta}\sin\hat{\phi}}{\hat{n}_4 f_y} d_v}{\sqrt{d_u^2 + d_v^2}\sqrt{(\frac{\sin\hat{\theta}\cos\hat{\phi}}{\hat{n}_4 f_x})^2 + (\frac{\sin\hat{\theta}\sin\hat{\phi}}{\hat{n}_4 f_y})^2}} \\
&= \frac{\frac{\sin\hat{\theta}}{\hat{n}_4}(\frac{\cos\hat{\phi}}{f_x} d_u + \frac{\sin\hat{\phi}}{f_y} d_v)}{\frac{\sin\hat{\theta}}{\hat{n}_4}\sqrt{d_u^2 + d_v^2}\sqrt{(\frac{\cos\hat{\phi}}{f_x})^2 + (\frac{\sin\hat{\phi}}{f_y})^2}} \\
&\approx \frac{(\cos\hat{\phi} d_u + \sin\hat{\phi} d_v)}{\sqrt{d_u^2 + d_v^2}\sqrt{(\cos\hat{\phi})^2 + (\sin\hat{\phi})^2}} \qquad \text{with } f_x \approx f_y \text{ for most common cameras} \\
&= \frac{\cos\hat{\phi} d_u + \sin\hat{\phi} d_v}{\sqrt{d_u^2 + d_v^2}} \qquad \text{with } (\cos\hat{\phi})^2 + (\sin\hat{\phi})^2 = 1
\end{aligned}
$$

According to the above formulation, in most cases, CPD loss persistently guides $\hat{\phi}$, while $\hat{\theta}$ and $\hat{n}_4$ still degenerate. The characteristic of CPD loss implies two underlying facts. One is that CPD loss facilitates the optimization and convergence on $\hat{\phi}$, so that training process becomes friendly. The other one is that CPD loss is not able to counter the polar angle collapse, so APR is crucial to recover the polar angle estimates from the collapse.

D. More Analysis on CPR loss

D.1. The absolute manner of CPR loss

Recall below Eq. (7) that is transformed from Eq. (6). To derive Eq. (8) from it, the term C has been discarded by the first order partial derivatives of u and v. It means that Eq. (8) (CPD loss) guides estimation in a relative manner, which could not absolutely satisfy Eq. (6) (CPR loss). There usually exists offsets defined by term C between absolute and relative estimates.

$$
Z^{-1}(u,v) = \frac{\sin\theta\cos\phi}{n_4 f_x} u + \frac{\sin\theta\sin\phi}{n_4 f_y} v + \underbrace{(\frac{\cos\theta}{n_4} - \frac{\sin\theta\cos\phi}{n_4 f_x} c_x - \frac{\sin\theta\sin\phi}{n_4 f_y} c_y)}_{\text{This term } C \text{ indicates the offset between absolute and relative estimate.}}
$$

Meanwhile, CPD loss hardly contributes to regulate $\hat{\theta}$ as discussed in Appendix C.2. Thereby, even though $\hat{\phi}$ is constrained by CPD loss and $\hat{\theta}$ is regularized by APR, the surface normal estimation could be still relative and biased. In order to counter this, our CPR loss derived from Eq. (6) plays its critical role. Along with CPD loss that persistently constrains $\hat{\phi}$ and APR that recovers the $\hat{\theta}$ from collapse, CPR loss compensates the absolute relationship by further adjusting estimation. In details, the biases on $\hat{\theta}$ from APR are suppressed by CPR loss (see next Appendix D.2). At the same time, CPR loss leverages $\hat{n}_4$ in a smooth L1 manner, which is more robust to noises and discontinuity. Eventually, the surface normal estimation is rectified and polished, which has been quantitatively analyzed and quantitatively visualized in our main manuscript (see ablation studies in our main manuscript).

D.2. APR bias suppression by CPR loss

As shown in Fig. 1b, APR has stable minimums around $\pm\pi/2$, which could incur bias on $\hat{\theta}$ at these minimums. Without special treatment on $\hat{\theta}$, CPD loss with APR is confronted with this biased situation, which is presented in Fig. 3 (see the red curve). The density of polar angle estimates is abnormally accumulated around $\pm\pi/2$. However, ground truth is low on density near $\pm\pi/2$ (see the orange histogram).

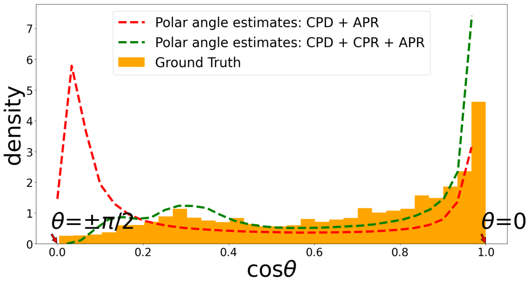


Figure 3: The APR bias suppression by CPR loss. The distributions are from the NYUv2 dataset as an example. The green and red curves are the envelopes of the distributions on polar angle estimates (expressed by $\cos\theta$) with and without CPR respectively. The orange histogram is from the ground truth. CPR loss effectively rectifies the biased distribution.

The absolute manner of CPR loss contributes to suppress the polar angle bias from APR (see the green curve in Fig. 3). The reason is that the majority of polar angle estimates near $\pm\pi/2$ contradicts the absolute constraint defined by CPR loss. Specifically, 3D scenes with $\theta = \pm\pi/2$ describe the invisible planes parallel to the vision sight, so the scenes with $\theta \approx \pm\pi/2$ (the APR minimums) should be rare in common real world scenes because of visibility. This prior is also the reason why we adjust APR minimums to near $\pm\pi/2$ via power of $1/4$. When more polar angle estimates are skewed to $\approx \pm\pi/2$ by APR, the reasonable depth map (low density on $\hat{\theta} \approx \pm\pi/2$) becomes more inconsistent to these biased surface normals (high density on $\hat{\theta} \approx \pm\pi/2$), which leads to the increasing of CPR loss. This contradiction prevents APR to skew the polar angle estimates, and it is clearly observed from the training loss, as shown in Fig. 4.

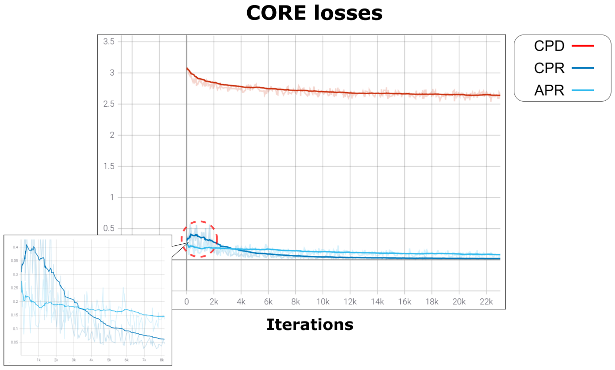


Figure 4: The curves of CORE losses during training. CPR loss and APR are obviously adversarial at early training stage, where the dramatic dropping on APR greatly increases CPR loss.

E. KITTI Evaluation conditioned on distance

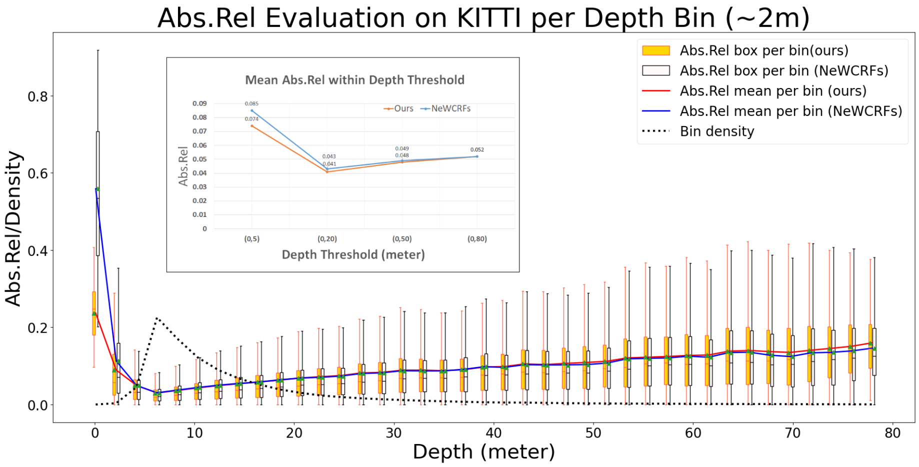


Figure 5: The Abs.Rel conditioned on distance. The green triangle indicates mean. The subplot shows the mean Abs.Rel within a depth threshold.

We conduct KITTI evaluation conditioned with distance in Fig. 5. Each box plot represents statistics of pixel Abs.Rel falling in the depth bin. The mean Abs.Rel per bin and within a depth threshold (subplot) both indicate evidently that our design enhances the depth estimation for near scenes ($<$ 20m), especially for closer scenes ($<$ 5m). From box plots, we can observe more statistic details such as max error and median, which are mostly ameliorated for near scenes as well. When distance $>$ 50m, our Abs.Rel becomes slightly worse. However, the depth density in this range is low, and the overall impact is insignificant.

F. Geometry Consistency

As shown in Fig. 6, we spot that the depth estimates by our method have more similar distribution of density to the ground truth, including subtle trend of the envelope (marked by red markers). This fact implies that our method can distinguish more consistent geometry to the ground truth from the scenes, which is similarly demonstrated by the point cloud visualization in our main manuscript. The reason lies in that our CORE losses mutually refine the depth the surface normals estimation for better geometry information.

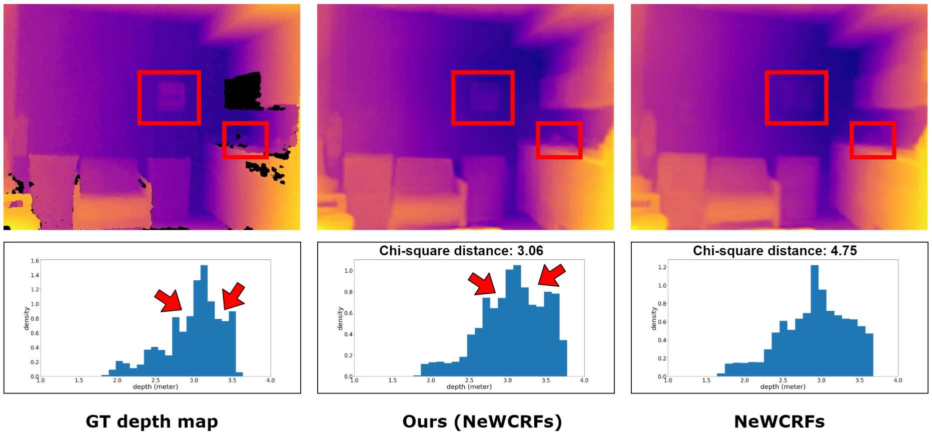


Figure 6: Illustration of geometry consistency. Bottom row shows according histograms for the top. Our depth estimation distinguishes more objects marked with red boxes, and the histogram has smaller Chi-square distance (on the top of the histogram) to the ground truth.

G. More qualitative results

In the main manuscript, we demonstrated some qualitative results with NeWCRFs [7] enhanced by our method. Here, we present additional qualitative results and comparison for BTS [5] and Adabins [1]. Also, various datasets [6, 3, 2] are involved to demonstrate the effectiveness of our method, including depth and surface normal estimation.

G.1. NYUv2

With Fig. 7, we present the qualitative results on NYUv2 for BTS [5] and Adabins [1], and compare the results between them and that enhanced by our method. Additionally, Fig. 8 illustrates more qualitative results of surface normal estimation by our method.

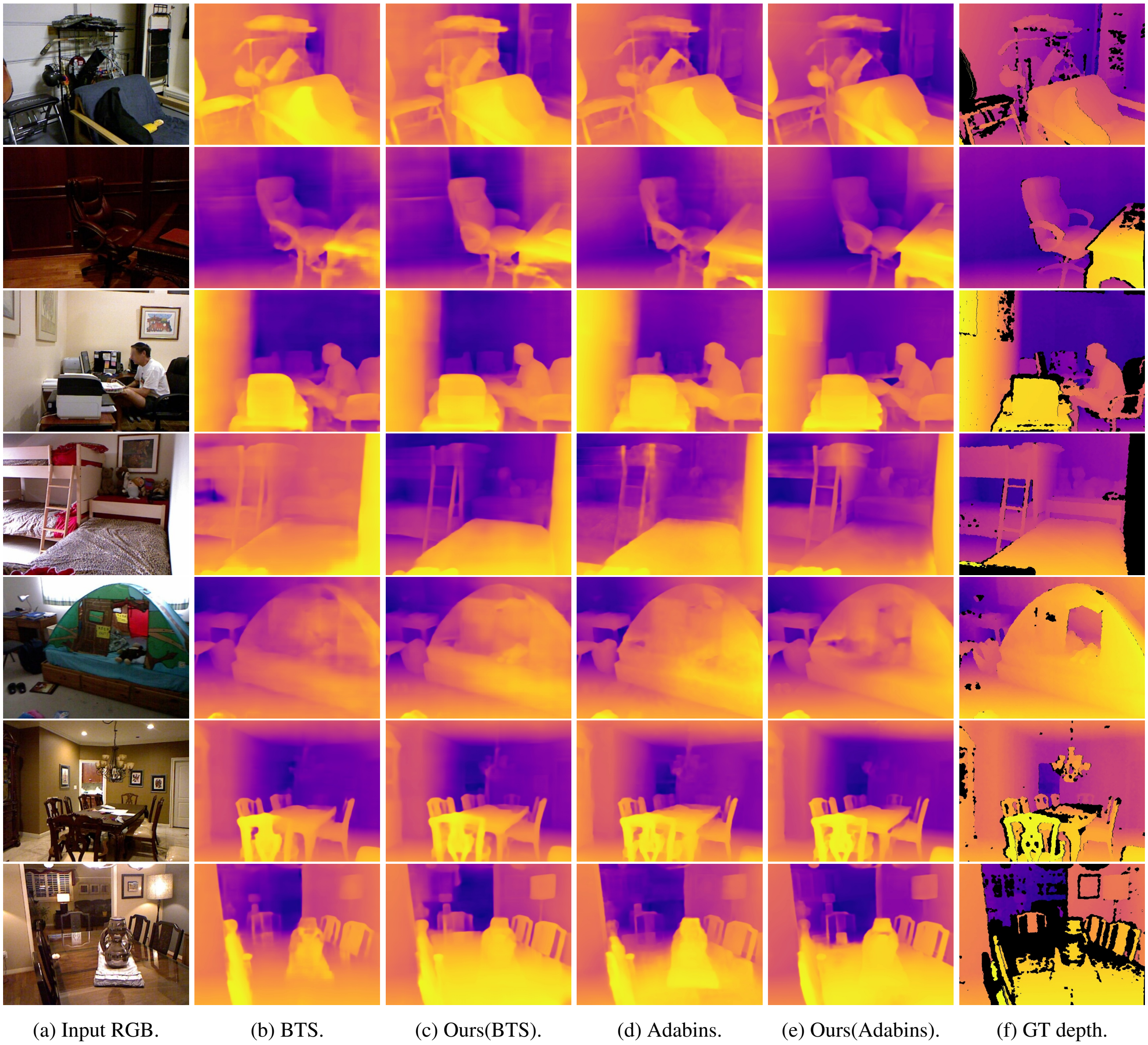
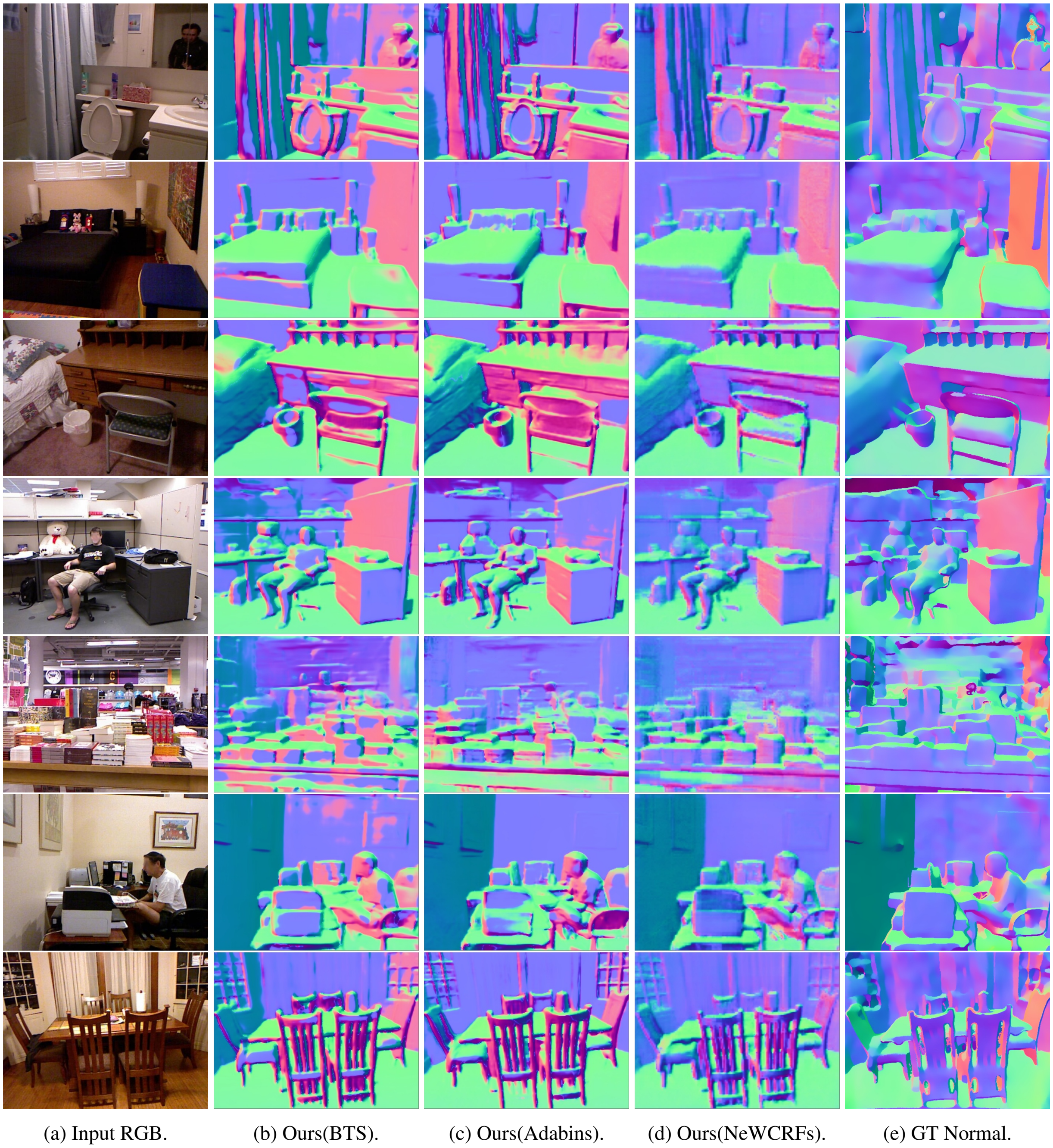


Figure 7: More qualitative depth estimation results on NYUv2 for BTS [5] and Adabins [1]. Our method rectifies geometry details and occlusions for the original approaches, and provides more consistent depth map to the ground truth. Zoom-in and best view in color.

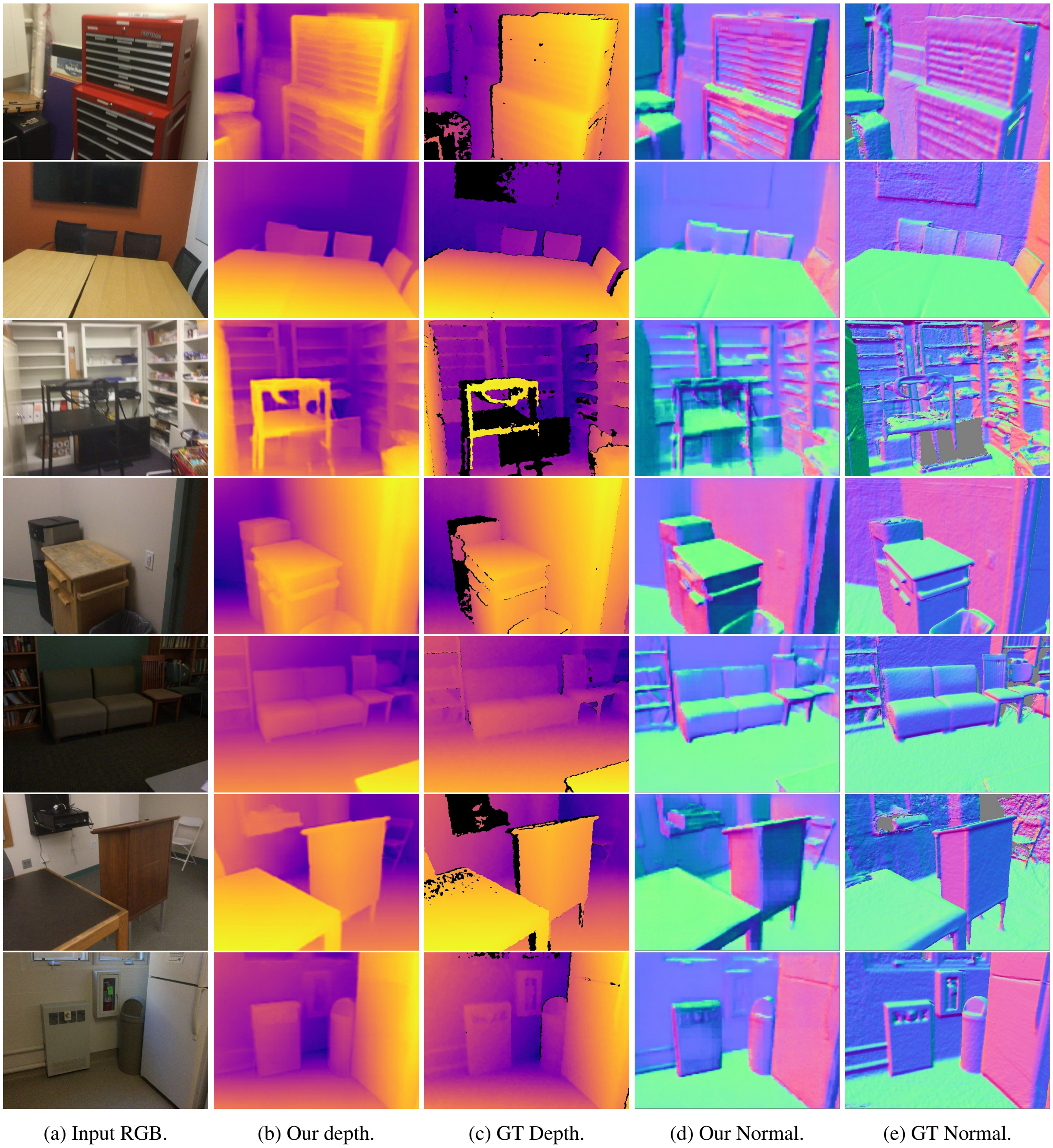


(a) Input RGB. (b) Ours(BTS). (c) Ours(Adabins). (d) Ours(NeWCRFs). (e) GT Normal.

Figure 8: Qualitative results on the weakly-supervised surface normal prediction. NeWCRFs [7] with our method indicates less noises for surface normal estimation. Our method enables BTS [5], Adabins [1] and NewCRFs [7] to gain the capability of surface normal prediction. Zoom-in and best view in color.

G.2. ScanNetv2 (cross-evaluation)

We employ Fig. 9 to demonstrate some qualitative results on ScanNetv2 for both depth and surface normal prediction. The results are predicted by our model well trained on NYUv2 without any fine-tuning. Ground-truth surface normals in Fig. 9 are generated by FrameNet [4].



(a) Input RGB. (b) Our depth. (c) GT Depth. (d) Our Normal. (e) GT Normal.

Figure 9: Qualitative results on ScanNetv2 (cross-evaluation) with NeWCRFs [7] enhanced by our method. The ground truth surface normals are generated by [4]. Zoom-in and best view in color.

G.3. KITTI

As presented in our main manuscript, improvement of depth estimation by our method on KITTI are not as significant as that on NYUv2. However, some geometry rectifications are still observed in Fig. 10. Along with depth estimation, in Fig. 11, we demonstrate the qualitative results of surface normal prediction for KITTI as well.

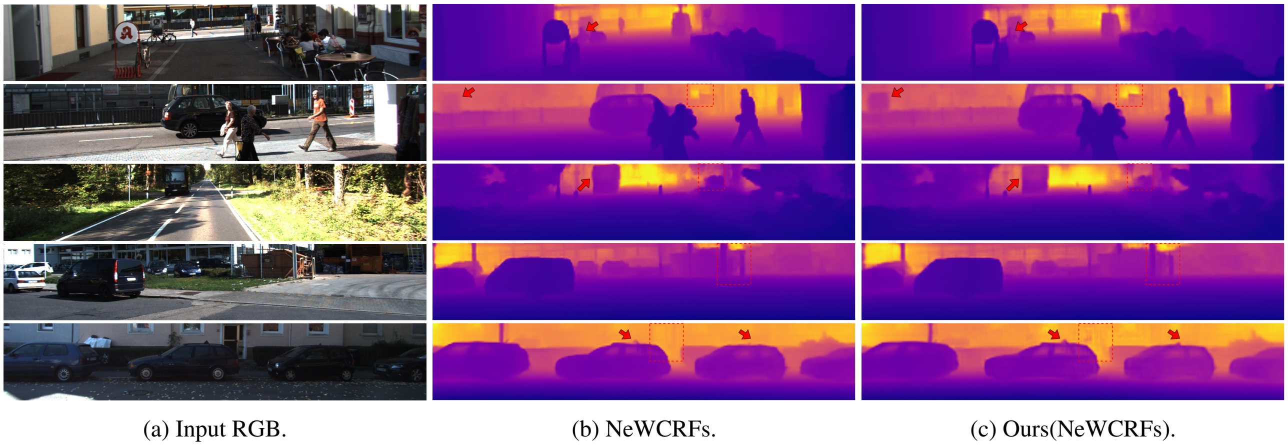


Figure 10: Qualitative depth estimation on KITTI with NeWCRFs [7] and our method. Some geometry details, especially the occlusion between objects, have been rectified by our method, as indicated by red markers and boxes. The co-planar information helps to distinguish the belonging, when two objects have occlusion in space. Zoom-in and best view in color.

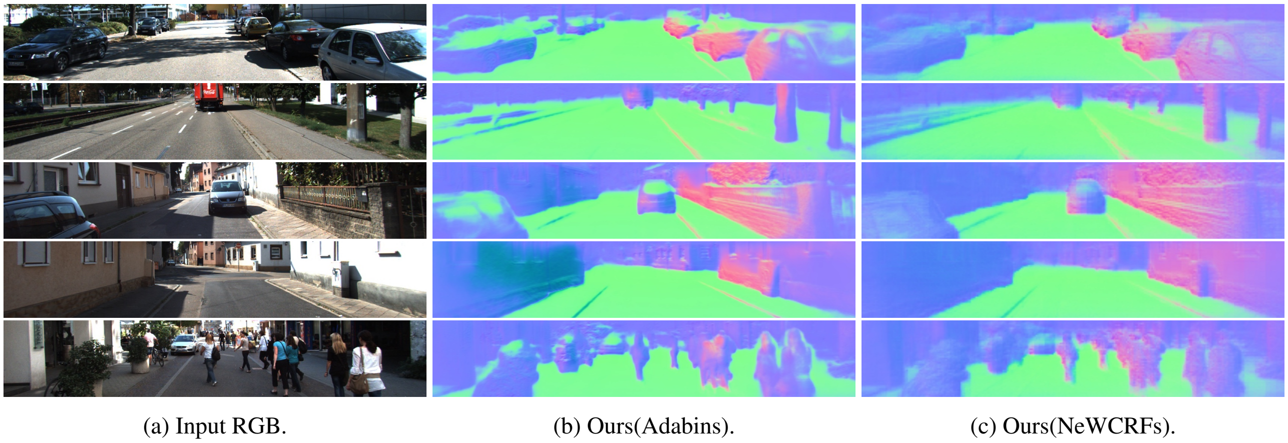
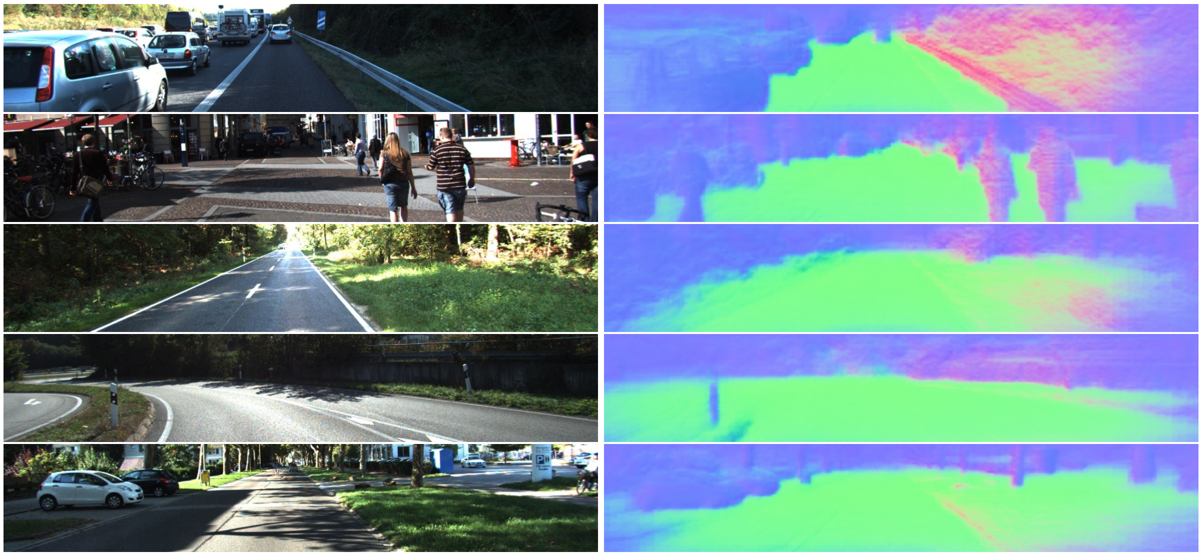


Figure 11: Qualitative surface normal predictions on KITTI. The NeWCRFs enhanced by our method provides less noisy surface normal prediction. Zoom-in and best view in color.

H. Failure cases

Our method has some difficulties to make a decent surface normal prediction in some outdoor scenes (*e.g.*, KITTI dataset) as shown in Fig. 12. Surface normal estimates for KITTI are not able to provide as many details as the prediction results on NYUv2 and ScanNetv2 that are indoor scenes.

We speculate a few reasons for the phenomenon and failure. Firstly, the sparse depth map provided by KITTI could be a reason for the lack of geometry details. Secondly, distortions introduced by the camera with large Field of View (FoV) incur noisy surface normal prediction near image borders and corners. Thirdly, the co-planar assumption is not restrictively established in some outdoor scenes, especially for the crowds and woods at far distance. All these will be investigated in the future work.



(a) Input RGB.

(b) Ours(NeWCRFs).

Figure 12: Failure cases of surface normal prediction on KITTI. In some outdoor scenes, our surface normal prediction would have difficulties to distinguish geometry details. Zoom-in and best view in color.

References

[1] 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. 6, 7

[2] 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. 6

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

[4] 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. 8

[5] 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. 6, 7

[6] 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. 3, 6

[7] 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. 6, 7, 8, 9