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

Supplementary Material

A. Overview

This document supplements the method details and additional experimental results. In Section B, we discuss the structured plane generation module in detail. In Section C, we report the results of integrating our idea with recent SOTA frameworks. In Section D, we provide more ablation study. In Section E, we present the complexity of the model and the speed of inference. In Section F, we provide more qualitative results.

B. Details of Structured Plane Generation Module

More details of our proposed structured plane generation module is shown in Fig. B1. Given the predicted planar normal maps, planar offset maps, and depth maps of two adjacent views, we generate the spatial geometric cues (i.e., $\bar{\boldsymbol{N}}_c$ and $\bar{\boldsymbol{O}}_c$) and temporal geometric cues (i.e., $\boldsymbol{N}_{n2c}$, $\boldsymbol{O}'_n$ and $\boldsymbol{O}''_n$). In order to recover the correct plane representation, this module first uses spatial-temporal geometric cues to guide two planar properties to recover the approximate scene structure. Then, the two planar properties are jointly optimized based on the planar uniqueness principle to ensure that coplanar points recover a unified representation.

C. Integrating Our Idea with SOTA Framework

Table C1 reports the results of integrating our idea with recent SOTA frameworks on KITTI, including CADepth-Net, RA-Depth and MonoViT. The results clearly indicate that our method consistently outperforms these frameworks across various backbones, showcasing its robustness and generalizability. **When combined with full RA-Depth model, our method sets a new SOTA**, highlighting the effectiveness of our method in improving depth estimation and advancing current model performance.

D. More Ablation Study

Table C2 presents the detailed ablation results of the structured plane generation module, with each component added to the baseline model separately. The results show that every design choice positively impacts overall performance.

E. Model Complexity and Speed Evaluation

Table D3 reports the parameter complexity (#Params), computation complexity (GLOPs), and inference speed on the KITTI [1] dataset. We perform inference at a resolution of 640 × 1920, and set the batch size to 16. The models for all comparison methods were inferred on the same platform with NVIDIA RTX 3090 GPU. As can be seen from this table, our model has a similar number of parameters as most existing depth estimation methods, such as Monodepth2-R18 [2], Dynamo-Depth [3], which allows it to be used on edge devices. Moreover, our model achieve similar computation complexity and inference speed as the existing lightweight model Lite-Mono [5]. Therefore, our proposed GeoDepth is able to be practically applied.

Method	Backbone	Sq Rel↓	RMSE ↓	δ<1.25↑
CADepth-Net	ResNet50	0.769	4.535	0.892
GeoDepth	ResNet50	**0.745**	**4.478**	**0.896**
RA-Depth	HRNet18	0.632	4.216	0.903
GeoDepth	HRNet18	**0.624**	**4.169**	**0.904**
MonoViT	MPViT	0.708	4.372	0.900
GeoDepth	MPViT	**0.662**	**4.237**	**0.902**

Table C1. Like-for-like comparisons.

Method	RMSE ↓	Sq Rel↓	δ<1.25↑
Baseline+P2D	4.436	0.740	0.896
+Spatial	4.431	0.738	0.896
+Temporal	4.428	0.723	0.896
+Uniqueness	4.416	0.724	0.896

Table C2. **Detailed ablation of structured plane generation**. Baseline: Predict depth directly from a single image. P2D: Using Plane-to-Depth modeling. Spatial: Use planar normal and offset alignment on spatial dimensions. Temporal: Use planar normal and offset alignment on temporal dimensions. Uniqueness: Using planar uniqueness alignment.

F. More Qualitative Results

To gain a clear understanding of the explicit planar representations we explore to generate accurate and continuous depth maps, we present more qualitative results in Fig. E2, including planar offsets, planar normals, and depth maps. Observing the second and third columns reveals that pixels in the same plane share unique representations, indicating that our method models coplanar points as unified planar representations rather than discrete points. Combined with our plane-to-depth modeling, the depth variations of coplanar points are exclusively linked to pixel coordinates.

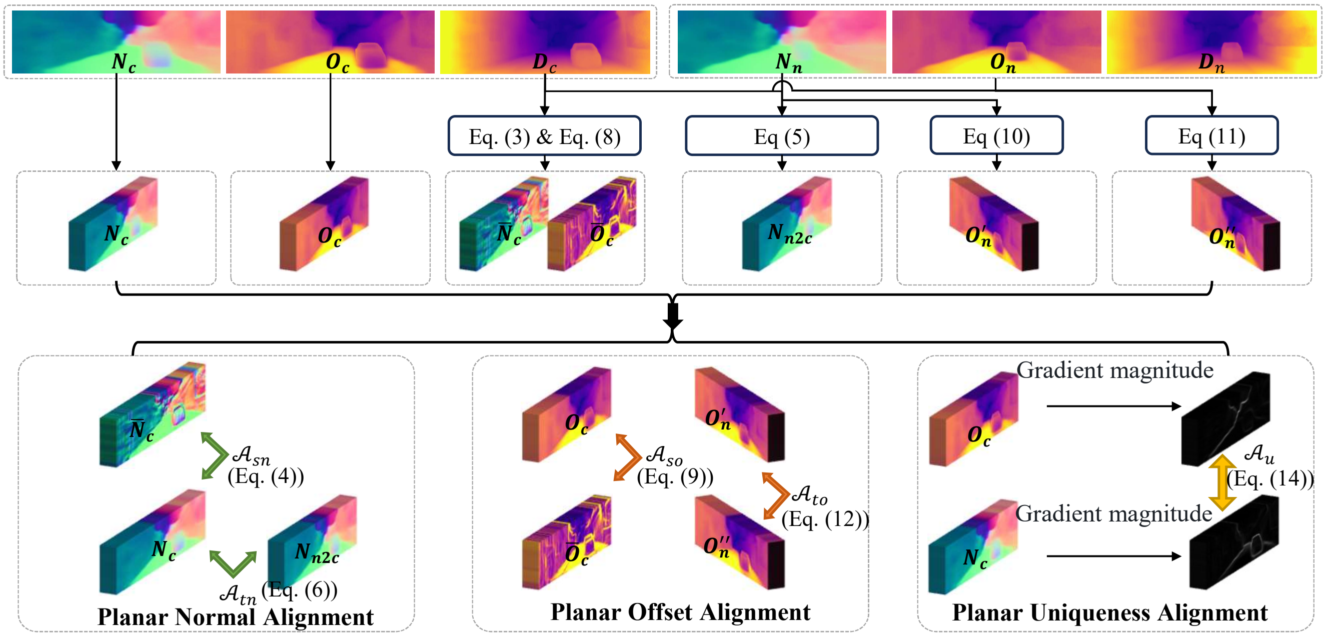


Figure B1. **Details of structured plane generation module**.

Method	#Params	GFLOPs	Speed	RMSE ↓	RMSE log ↓	Sq Rel ↓	Abs Rel ↓
GeoNet [4]	31.6M	-	-	5.567	0.226	1.060	0.149
Monodepth2-R18 [2]	14.3M	8.04G	4.0ms	4.863	0.193	0.903	0.115
Monodepth2-R50 [2]	32.5M	16.7G	5.3ms	4.642	0.187	0.831	0.110
Lite-Mono [5]	3.07 M	5.03G	4.9ms	4.561	0.183	0.765	0.107
Dynamo-Depth (MD2) [3]	14.3M	8.04G	6.2ms	4.850	0.195	0.864	0.120
Dynamo-Depth [3]	8.77M	11.2G	10.1ms	4.505	0.183	0.758	0.112
GeoDepth	10.0M	11.9G	4.2ms	**4.381**	**0.176**	**0.694**	**0.100**

Table D3. **Model complexity and speed evaluation**. We compare parameters (#Params), giga floating-point operations per second (GFLOPS), and inference speed on the KITTI [1] dataset. The input size is 640 × 192, and the batch size is 16. All models are inferred on the same platform with NVIDIA RTX 3090 GPU. “-” indicates that the method is not open source code and we cannot make inferences.

Therefore, by restoring the structured representation among pixels, we are able to mitigate depth discontinuities and generate accurate and continuous depth maps.

References

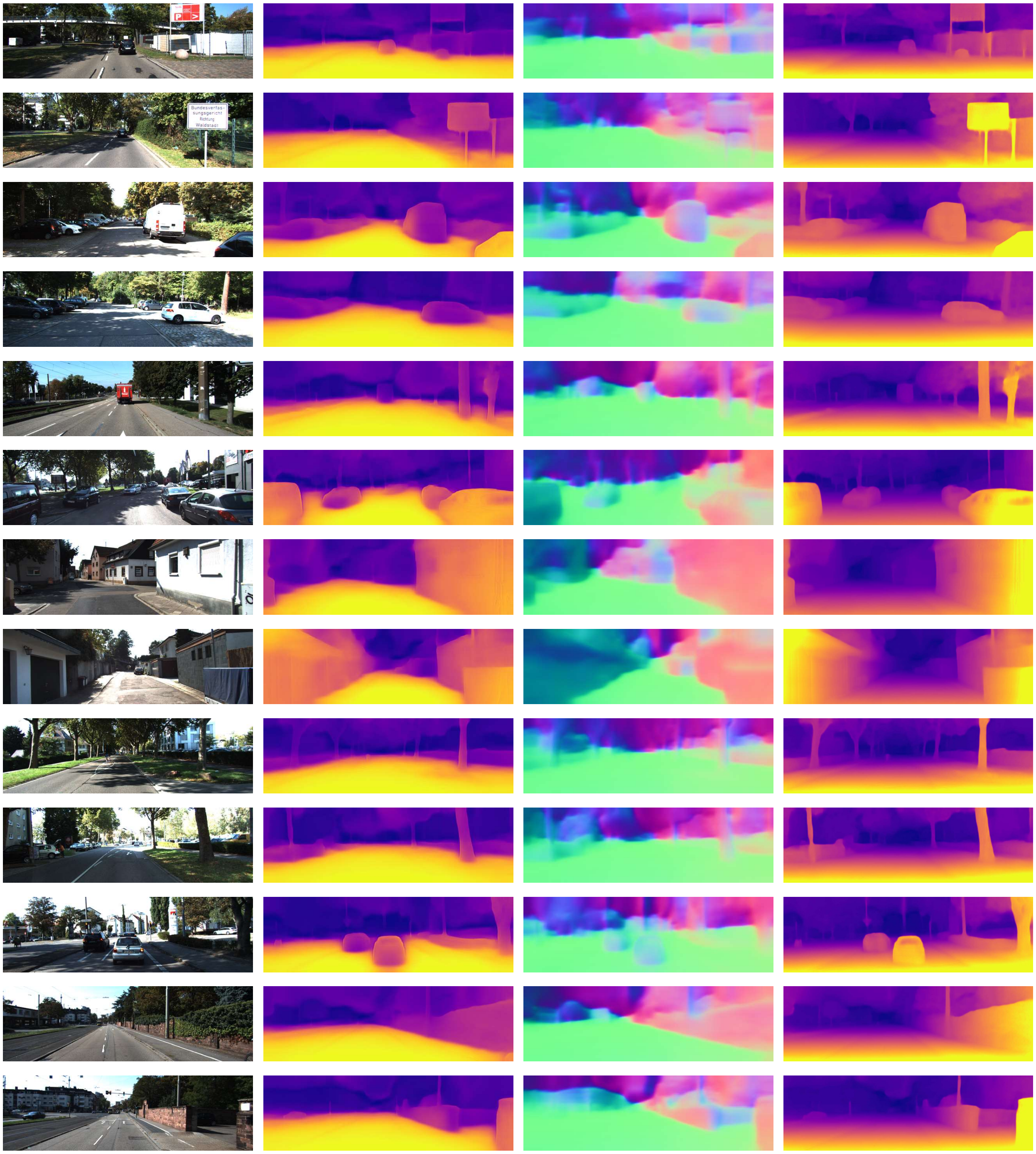
[1] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 3

[2] 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, 2

[3] Yihong Sun and Bharath Hariharan. Dynamo-depth: Fixing unsupervised depth estimation for dynamical scenes. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[4] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1983–1992, 2018. 2

[5] 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, 2



(a) Input (b) Planar offset (c) Planar normal (d) Depth

Figure E2. More qualitative results on the KITTI dataset [1].