Supplementary Material for
Depth Prompting for Sensor-Agnostic Depth Estimation

A. Zero-shot Inference on Commercial Sensors.

VOID [6] provides RGB-D images with a resolution of 640×480, captured using a RealSense D435i. We utilize sparse depth images with about 150, 500, and 1500 depth points as input, which are estimated feature points by the SLAM system, VIO [3]. We evaluate the adaptability of both our method and other SoTA approaches across a range of density levels to assess their versatility, and conduct a quantitative experiment on the test set provided by [6]. As shown in Fig. A and Tab. A, while other methods struggle to infer depth maps, our method consistently shows comparable depth map quality, which indicates the efficacy of our approach as a solution for sensor bias problems.

SUN RGB-D [5] contains RGB-D images from four different sensors, offering a diverse range of scenes and sensors, e.g., Intel RealSense 3D Camera, Asus Xtion LIVE PRO, Microsoft Kinect V1 and V2. The input images are resized to 320×240, and the center cropped to 304×228. The total of 1000 scenes, where each sequence is roughly 20 seconds long and annotated every 0.5 seconds, is officially split into train/val/test set with 700/150/150 scenes. Using this dataset, we demonstrate the versatility of our method across different sensors. According to Fig. B and Tab. A, other SoTA methods suffer from reconstructing depth maps in areas where the raw depth data is missing. In contrast, our method effectively compensates for the missing parts.

Apple ARKit [4] is Apple's Augmented Reality (AR) development platform for iOS mobile devices. The Apple iPhone/iPad Pro devices use depth derived from RGB to recover scene details lost in the sparse LiDAR measurements. It provides high-resolution RGB images at 1440×1920 pixels, low-resolution depth maps at 192×256 pixels, and the corresponding confidence map. We compare the depth map output from the ARkit and the inference from our method using the sparse depth. For fair comparison, we first mask out unreliable depth information with a lower confidence score, and then reconstruct scene depths with the same resolution of ARKit. Fig. C showcases that our method achieves better results than ARKit results. We suggest that the limited sensor range of Apple's LiDAR often leads to bad predictions, typically missing parts of the far region. However, thanks to the foundation model integrated into our framework, our approach successfully overcomes this limitation, yielding high-quality depth maps even in distant areas.

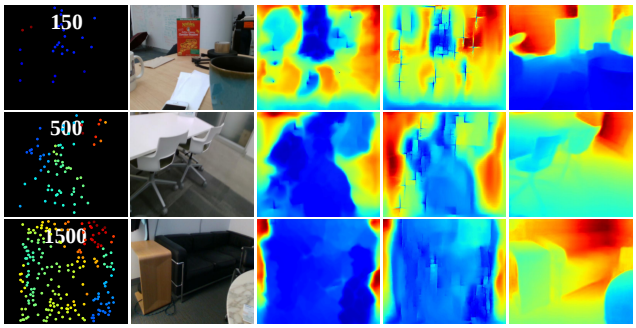

Figure A. Zero-shot inference on VOID dataset. Since it offers various inputs consisting of 1500, 500, and 150 feature points, we can validate our method's robustness in handling sensor biases.

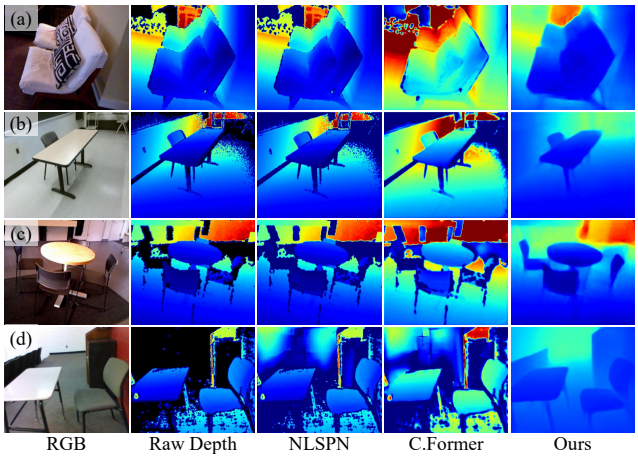

Figure B. Zero-shot inference on SUN-RGBD consisting of various sensor types, e.g., Microsoft Kinect (a),(b) V1&V2, (c) Asus Xtion, and (d) Intel Realsense.

nuScenes [1] offers detailed point clouds captured from a 32-Line LiDAR, along with high-resolution images of 1600×900 pixels from six cameras on the vehicle. For inference using models pretrained on the KITTI dataset [2], we first perform first center cropping of the images to a size of 352×1600 pixels, and then resizing them to 240×1216 pixels. As shown in Fig. D, our method constructs high-fidelity depth maps in the new sensor across both day and night conditions. Notably, our model, which is initially trained on 64-Line Velodyne LiDAR, shows exceptional performances when an unseen LiDAR sensor is applied at inference time.

	VOID 150			VOID 500			VOID 1500			SUN-RGBD		
	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1
CSPN	1.0747	0.7991	0.2745	0.7695	0.4776	0.4730	0.4046	0.2121	0.7286	0.5552	0.2464	0.8468
NLSPN	0.6964	0.5063	0.3572	0.4672	0.2853	0.5618	0.2563	0.1339	0.7801	0.5910	0.2665	0.8441
CompletionFormer	0.8445	0.6976	0.2030	0.6090	0.4447	0.3541	0.3430	0.2158	0.5794	0.9615	0.8574	0.0564
Ours	**0.1955**	**0.1358**	**0.7243**	**0.1637**	**0.1031**	**0.7457**	**0.1580**	**0.098**	**0.7582**	**0.3962**	**0.2710**	**0.8402**

Table A. Quantitative results on VOID and SUN-RGBD dataset.

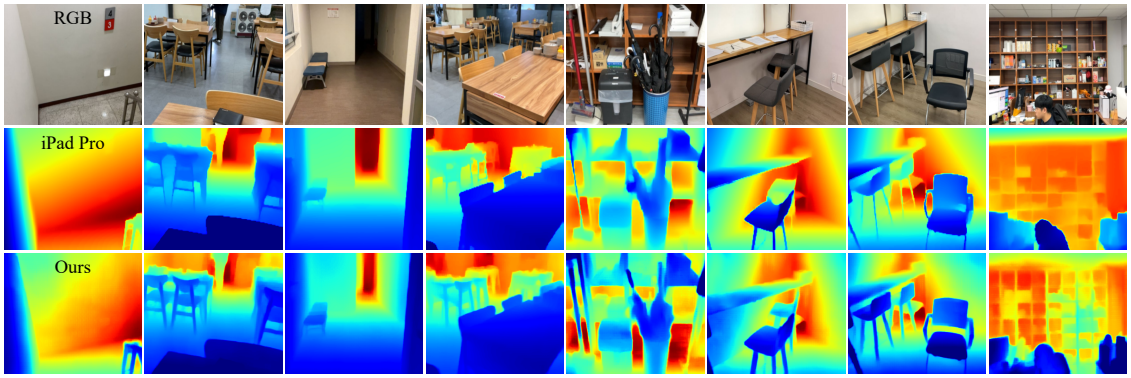

Figure C. Zero-shot inference on Apple LiDAR. Due to the scanning range of Apple Lidar, the results have limitations in representing the depth map for distant regions. Our method mitigates this issue by leveraging the knowledge from our foundation model.

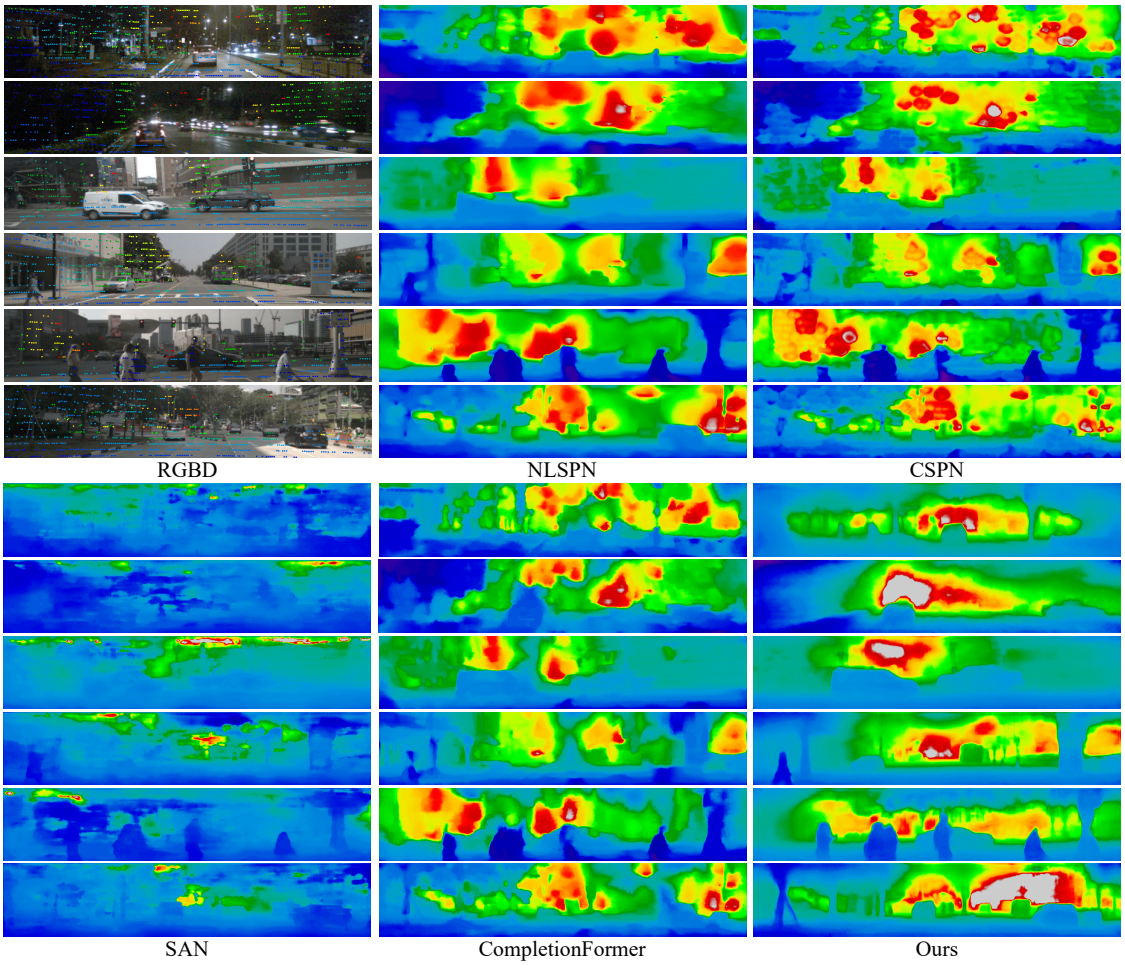

Figure D. Zero-shot inference on nuScenes dataset. We test the versatility of the proposed method in other outdoor datasets. Despite sensor variations, our method demonstrated stable generalization performance, unlike other SoTA methods that suffer from sensor bias issues.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

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

[3] Eagle S Jones and Stefano Soatto. Visual-inertial navigation, mapping and localization: A scalable real-time causal approach. *The International Journal of Robotics Research (IJRR)*, 30(4): 407–430, 2011. 1

[4] Zainab Oufqir, Abdellatif El Abderrahmani, and Khalid Satori. Arkit and arcore in serve to augmented reality. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–7. IEEE, 2020. 1

[5] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[6] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters*, 5(2):1899–1906, 2020. 1