Depth Prompting for Sensor-Agnostic Depth Estimation

Jin-Hwi Park[1], Chanhwi Jeong[1], Junoh Lee[2] and Hae-Gon Jeon[1,2*]
[1]AI Graduate School, [2]School of Electrical Engineering and Computer Science, GIST, South Korea
{jinhwipark,chanhwij,juno}@gm.gist.ac.kr, haegonj@gist.ac.kr

Abstract

Dense depth maps have been used as a key element of visual perception tasks. There have been tremendous efforts to enhance the depth quality, ranging from optimization-based to learning-based methods. Despite the remarkable progress for a long time, their applicability in the real world is limited due to systematic measurement biases such as density, sensing pattern, and scan range. It is well-known that the biases make it difficult for these methods to achieve their generalization. We observe that learning a joint representation for input modalities (e.g., images and depth), which most recent methods adopt, is sensitive to the biases. In this work, we disentangle those modalities to mitigate the biases with prompt engineering. For this, we design a novel depth prompt module to allow the desirable feature representation according to new depth distributions from either sensor types or scene configurations. Our depth prompt can be embedded into foundation models for monocular depth estimation. Through this embedding process, our method helps the pretrained model to be free from restraint of depth scan range and to provide absolute scale depth maps. We demonstrate the effectiveness of our method through extensive evaluations. Source code is publicly available at https://github.com/JinhwiPark/DepthPrompting.

1. Introduction

Scene depths have been used as one of the key elements for various visual perception tasks such as 3D object detection [65], action recognition [66], and augmented reality [23, 75], etc. For accurate depth acquisition, there have been various attempts in the computer vision field. Since the advances in deep learning, its powerful representational capacity has been applied to explain scene configurations, which is feasible even with only single images. Unfortunately, single image depth estimation cannot produce metric scale 3D depths when camera parameters change and out-of-distributions on unseen datasets happen [20].

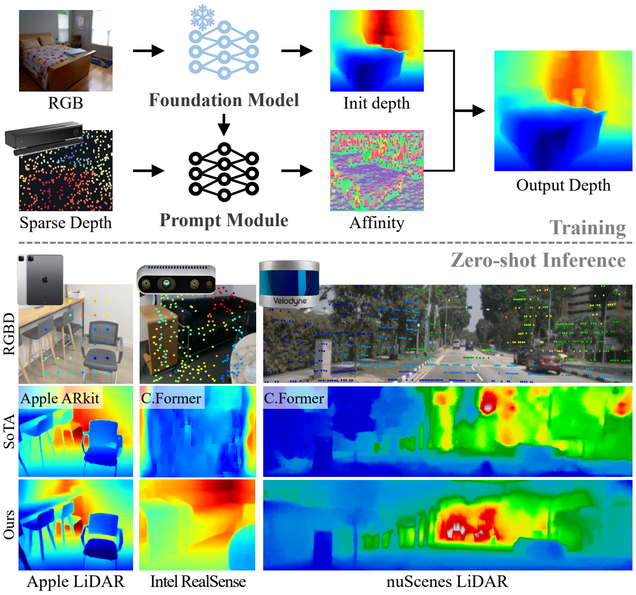


Figure 1. An overview of our depth prompting for sensor-agnostic depth estimation. Leveraging a foundation model for monocular depth estimation, our framework produces a high-fidelity depth map in metric scale and provides impressive zero/few-shot generality. C.Former indicates CompletionFormer [73]. More details and examples are reported in Sec. 4.5 and supplementary materials.

Toward depth information easy to acquire in metric scale, active sensing methods such as LiDAR (Light Detection and Ranging) [54], ToF (Time of Flight) [30], and structured light [72] have gained interest as a practical solution. Although the active sensing methods enable real-time scene depth acquisitions in a single shot, they only provide sparse measurements. For dense predictions, spatial propagation, modeling an affinity among input image pixels, is necessary [9, 35, 36, 45, 46, 73]. Note that its affinity map is constructed based on input images, and is jointly optimized with fixed depth patterns. Different from real-world scenarios where various types of depth sensors (e.g., Velodyne LiDAR [54], Microsoft Kinect [74], Intel RealSense [27], and Apple LiDAR sensor [40], etc.) are used, the mainstream of standard benchmarks [59] for this task is to only use KITTI dataset [17] captured from a 64-Line LiDAR and random

[*]Corresponding author

samples from Kinect depth camera in NYUv2 dataset [55].

In this work, our primary goal is to build a sensor-agnostic depth estimation model that faithfully works on the various active depth sensors. Inspired by pioneer works in visual prompt methods like SAM [29], we design a novel depth prompt module used with pre-trained models for monocular depth estimation. The depth prompt first encodes the sparse depth information and then fuses it with image features to construct a pixel-wise affinity. A final refinement process is performed with both the affinity and an initial depth map from the pre-trained depth models. To take full advantage of pre-trained models, we conduct a bias tuning [7], which is a well-known memory-efficient technique when applied to pre-trained models. Our proposed method is fine-tuned for only 0.1% parameters of the models while keeping other parameters frozen.

Our key idea is to reinterpret prompt learning as spatial propagation. We aim to achieve an adaptive affinity from both the depth prompt and the knowledge of the pre-trained monocular depth model [34]. We demonstrate that the proposed method is generalized well to any sensor type, and it can be extended for various depth foundation models such as [4, 56], which are trained with large-scale datasets for monocular depth estimation to achieve relative scale prediction and zero-shot generalization. To do this, we utilize a variety of public datasets captured from off-the-shelf depth sensors and take real-world scenes by ourselves where the wide depth ranges and structures exist. Additionally, we conduct an extensive ablation study to verify the influence of each component within our framework and evaluate our methodology in a variety of real-world settings. This testing includes zero and few-shot inference exercises across different sensor types, further validating the robustness and adaptability of our proposed solution.

2. Related Works

2.1. Depth Estimation with Sparse Measurement

Accurate dense depth acquisition requires time-consuming and costly processes [32]. As a compromise between inference time and cost, sparse depth measurements based on active sensing manners have been considered as mainstream approaches. Leveraging the sparse measurements and its corresponding images, recent learning-based methods [9, 35, 36, 45, 46, 73] have been proposed to make dense predictions which enhance the depth resolutions same with the image resolutions via a spatial propagation process. However, due to the dependency on the specific sparse depth pattern and density according to the input devices, such models face challenges in real-world scenarios such as sensor blackouts [60], multipath interference [3, 44], and non-Lambertian surfaces [58], resulting in much fewer sample measurements.

Several studies have explored depth reconstruction from unevenly distributed and sparse input data [10, 19, 68]. However, they suffer from an issue on a range bias, which provides only limited scan ranges in training datasets. Works in [12, 67] focus on handling extremely sparse conditions (less than 0.1% over its input image); however, they fail to show the generalized performances on scenarios given relatively dense initial depth inputs. To address both issues, we adopt prompt engineering to achieve the model generalization, which has proven its powerful capacity by taking advantage of pre-trained models on downstream tasks, yet remains unsolved for depth-relevant tasks so far [6, 18, 29].

2.2. Prompt Engineering

Prompt engineering refers to designing specific templates that guide a model to complete missing information in a structured format (e.g., cloze set [6]), or generate a valid response along with given input (e.g., promptable segmentation [29]). With the recent success of the large language models [11, 53], works in [6, 14] demonstrate how various natural language processing (NLP) tasks can be reformulated to an incontext learning problem given a pre-defined prompt, which is a useful tool for solving the tasks and benchmarks [49, 50].

The advent of prompt engineering has been transformative, with some studies [29, 37, 51] extending its application to the computer vision field. Here, it is used to achieve a zero-shot generalization, enabling models to understand new visual concepts and data distributions that they are not explicitly trained on. The most relevant work [29] to this paper designs a promptable segmentation model. They construct a prompt encoder to represent user-defined points or boxes with a positional embedding [57]. We note that it is not a pixel-wise regression problem which is different from our task.

2.3. Foundation Model for Dense Prediction

Foundation models are designed to be adaptable for various downstream tasks by pretraining on broad data at scale [5]. In NLP field revolutionized by large-scale models such as GPT series [6, 49, 50], foundation models in the computer vision field have been becoming popular. Recent advancements in the large-scale foundation models with web-scale image databases have made significant breakthroughs, particularly involving image-to-text correspondences [24, 31, 51, 69]. These developments have paved the way for more efficient transfer learning [1, 51, 69] and make the zero-shot capabilities better [24, 31, 51].

Despite these advancements, the foundation models are primarily used in high-level vision tasks such as image recognition [24, 69], image captioning [1, 31, 69], and text-to-image generation [31]. When it comes to low-level vision tasks like depth predictions, these models do not seem to be suitable due to a lack of extensive image/depth data on a metric scale [33, 38]. To be specific, a web-scale dataset collection is infeasible because ground truth-level metric scale depths can be obtained only from sensor fusion manners [52]. Although some works [4, 56] have led to the creation of large and diverse datasets for monocular depth prediction, transferring the learned knowledge into other domains remains unexplored. To achieve the sensor-agnostic depth estimation regardless of scene configuration, we take fully advantage of the knowledge from the depth foundation model.

3. Sensor-agnostic Depth Estimation

In this section, we start by discussing the three biases which hinder sensor-agnostic depth prediction (Sec. 3.1). We then introduce the proposed depth prompt module. Here, we recast the depth prompt design as learning an adaptive affinity construction in spatial propagation for various types of sparse input measurements (Sec. 3.2). Lastly, we provide implementation details of the proposed module (Sec. 3.3).

3.1. Sensor Biases in Depth Estimation

Bias issues make learning-based models for visual perceptions hard to achieve their generality [2, 43]. There have been attempts to address the bias problems in image restoration [70], recognition [62] and generation [16]. Among them, the sensor-bias issue [13] is also considered as one of the crucial research topics. In particular, since a variety of depth sensors types is available, there is no generalization method to cover every depth sensor types, while the solutions to the same type (e.g., different LiDAR configurations [68]) exist. In this part, we empirically investigate 3-sensor biases, e.g., *sparsity*, *pattern*, and *range bias* before an introduction to our solution.

Firstly, if a learning-based model is trained on data with a certain density (e.g., 500 random samples in training), it will suffer from *sparsity bias*, which makes high-fidelity depth maps difficult if fewer samples are available in the test phase, as shown in Fig. 2-(b). The sparser sample measurements are also common due to sensor blackout, occlusion, or changing environmental conditions. This has hampered the practical utility of learning-based depth estimation in real-world scenarios. Second, *pattern bias* shows the performance degradation if depth patterns vary between training and test phases, even with the same number of depth points. When we intentionally shift the input depth pattern in the inference, it indicates that the existing model is biased toward the fixed location of input depth points in Fig. 2-(c). This makes a unified depth prediction model difficult to be applied to other sensor types. Lastly, *range bias*, arising when attempted to take scene structures beyond the limited scan range of the sensor, also prevents sensor-agnostic depth estimation as shown in Fig. 2-(d).

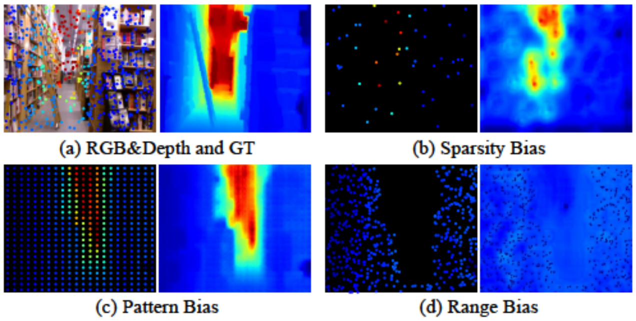

(a) RGB&Depth and GT
(b) Sparsity Bias
(c) Pattern Bias
(d) Range Bias

Figure 2. Examples of sensor biases. Depth estimation with an active sensor suffers from bias problems, including fixed density and pattern, and inherent scan range of sensors used.

3.2. Depth Prompting

To realize sensor-agnostic depth estimation without any sensor bias, we take an inspiration from prompt learning in NLP, which designs a specific template to guide a model for a valid response along with a given input [6, 14]. We aim to design a prompt module for depth modality by defining a unified embedding space to represent learned features from any type of input measurements (Fig. 3). Here, we use an input depth map as a template for our depth prompt module, and the sensor-agnostic depth prediction is achieved by fusing the template, features from the embedding space, and image features.

Revisiting Spatial Propagation. In this work, our key idea is to reinterpret the depth prompt design as spatial propagation, which predicts dense depth maps from input sparse measurements guided by image-dependent affinity weights. We formulate the conventional spatial propagation process:

$$D^{t+1}_{(x,y)} = A(x,y) \odot D^{0}_{(x,y)} + \sum_{(l,m)\in\mathcal{N}_{(x,y)}} A(l,m) \odot D^{t}_{(l,m)}, \quad (1)$$

where $D^{t}_{(x,y)} \in \mathbb{R}^{1\times H\times W}$ refers to a depth map for each propagation step t. (x,y) and H, W means a spatial coordinate and the height and width of an input image, respectively. $D^{0}_{(x,y)}$ and A indicate an initial depth and a pixel-wise affinity map, respectively. $\odot$ operator denotes an element-wise product. $(l,m) \in \mathcal{N}_{(x,y)}$ refers to 8-directional neighboring pixels over the reference pixel (x,y).

Even in the same scene, the affinity map A can vary according to the input depth type which is dependent on sensors used. That's, the affinity map should be adaptive to various input changes. However, affinity maps from previous spatial propagation methods [9, 35, 36, 45, 46, 73] are invariant because they are learned from a certain type of input depths. We address this issue by designing a depth encoder to learn features for a diverse set of sensors and by projecting them into the unified embedding space.

Depth Feature Extraction. To do this, we adopt an encoder-decoder structure to efficiently encode both positional and sparsity information of an input depth map. The encoder takes a depth map as an input, and then the decoder constructs an affinity map with the same size of the depth map. After that, the prompt embedding is combined with image features to bring boundary and context information.

To be specific, we use ResNet34 [21] to extract the depth features [39, 48, 71]. We downsample the features by $1/2$, $1/4$, $1/8$, $1/16$, and $1/32$, and then feed them into the decoder with skip connections. Given a sparse depth D_S , our depth prompt encoder $f_\mathcal{E}$ yields both a prompt embedding F^d and multi-scale features F^d_k as below:

$$F^d, F^d_k = f_\mathcal{E}(D_S), \tag{2}$$

where k is an index of the downsampled features.

Depth Foundation Model. Until now, large-scale depth models [4, 34, 56] have been tailored to monocular depth estimation. Leveraging a large-scale monocular depth dataset, the models are able to provide relative depth maps, which is the only option as a foundation model.

Given a single image $I \in \mathbb{R}^{3\times H\times W}$, the pre-trained depth model $f_\mathcal{F}$ outputs an initial depth map $\hat{D}_I$ and multi-scale intermediate features F^i_k:

$$\hat{D}_I, F^i_k = f_\mathcal{F}(I, \Theta_{f_\mathcal{F}}), \tag{3}$$

where $\Theta_{f_\mathcal{F}}$ indicates parameters of the foundation model, which keeps frozen during both training and inference.

Here, we need to effectively transfer the pre-trained knowledge of the foundation models to a range of sensors via prompt engineering. We adopt a bias tuning [7, 25], which is more effective for dense prediction tasks than other tuning protocols [22, 25]. This is used to update bias terms and to freeze the rest of the backbone parameters. As a result, the bias tuning contributes to preserving the high-resolution details and context information acquired in the initial extensive training phase [7], which will be analyzed in Sec. 4.4..

Next, to infer depth maps with absolute scales, we merge the relative depth from the foundation model with the sensor measurements. Due to the nature of spatial propagation, which mainly refines neighboring depth values over given seed points, we cannot obtain proper depth values for regions where no initial depth points are available. To address this limitation, we perform an additional processing with a least-square solver [42] to produce the consistent D_I when applied to other sensor types. The process uses both an initial depth from the foundation model and a sparse depth D_S in order to perform a global refinement. By solving the least square equation, we can obtain the solution $p \in \mathbb{R}$ as below:

$$\hat{p} = \min_p ||p\hat{D}^V_I - D_S||_F, \tag{4}$$

where $||\cdot||_F$ denotes the Frobenius norm, and $D^V_I \subset \hat{D}_I$ refers to a set of pixels corresponding to valid depth points D_S. The solution $\hat{p}$ is multiplied with the initial depth $\hat{D}_I$, whose result becomes $D^0_{(x,y)}$ of Eq. (1). Since this pre-calculation makes an initial depth for the spatial propagation in Eq. (1) better, we can cover larger unknown areas than those without Eq. (4).

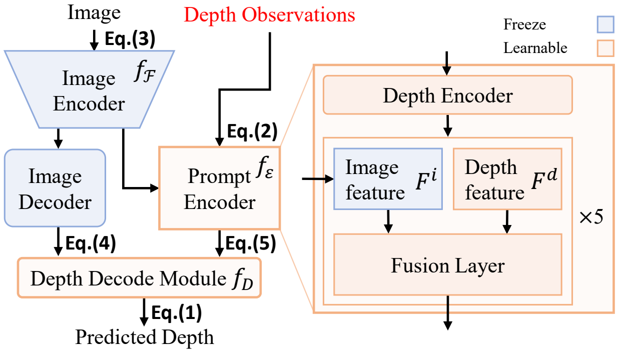


Figure 3. An overview of the proposed architecture. We design a depth prompt module to construct an adaptive affinity map A_{ada}, which guides the propagation of given depth information.

Decoder for Adaptive Affinity. A decoder in our prompt module reconstructs an affinity map using the image embedding from the foundation model encoder $f_\mathcal{F}$ (Eq. (3)) and a set of prompt embeddings from the prompt encoder $f_\mathcal{E}$ (Eq. (2)). We concatenate the prompt embeddings F^d, intermediate features F^d_i from the prompt encoder and multi-scale image features F^i_k, and then yield $A_{\text{ada}} \in \mathbb{R}^{C^2\times H\times W}$ where C is a hyper-parameter to define the propagation ranges and set to 7 as below:

$$A_{\text{ada}} = f_\mathcal{D}(F^d, F^d_k, F^i_k). \tag{5}$$

Finally, we substitute the conventional affinity map A in Eq. (1) into our A_{ada} above. Thanks to the feature fusion from both the prompt embedding and the foundation model with the bias-tuning, we can successfully decode an affinity map to account for different types of input measurements. In addition, the least square solver in Eq. (4) allows the spatial propagation to take consistent initial depth maps as input, regardless of the sensor variations. As a result, we achieve the adaptiveness/robustness in the proposed framework.

3.3. Implementation

Random Depth Augmentation. For more generality, we adopt random depth augmentation (RDA). We sample depth points from relatively dense depth maps to simulate sparser input depth scenarios. For example, we extract 4-Line depth values from 64-Line depth maps in the KITTI dataset [17]. In addition, we train our framework from general to extreme cases (e.g., from standard 500 random samples to only 1 depth point in the NYUv2 dataset [55])

Loss Functions. The proposed framework is trained in a supervised manner with the linear combination of two loss functions: (1) Scale-Invariant (SI) loss [15] for an initial depth from the depth foundation model $f_{\mathcal{F}}$ (Eq. (3)); (2) A combination of L_1 and L_2 losses for a final dense depth.

An initial depth map is predicted by minimizing the difference between $\hat{D}_I$ and its ground truth depth map D^{gt} for valid pixels $v \in V$. Let $\delta_v = \log \hat{D}_I(v) - \log D^{gt}(v)$, the SI loss L_{SI} is defined as below:

$$L_{\text{SI}}(\hat{D}_I, D^{gt}) = \frac{1}{|V|}\sum_{v\in V}(\delta_v)^2 - \frac{\lambda}{|V|^2}\left(\sum_{v\in V}\delta_v\right)^2, \quad (6)$$

where we set $\lambda = 0.85$ in all experiments, following the previous work [34].

Next, our framework infers a dense depth $\hat{D}$ in Eq. (1) based on the valid pixels $v \in V$ of its ground truth depth D^{gt} as well. For this, we use a loss L_{comb} based on both L_1 and L_2 distances as follows:

$$L_{\text{comb}}(\hat{D}, D^{gt}) = \frac{1}{|V|}\sum_{v\in V}\left(\left|\hat{D}(v) - D^{gt}(v)\right| + \left|\hat{D}(v) - D^{gt}(v)\right|^2\right), \quad (7)$$

In total, our framework is optimized by minimizing the final loss $\mathcal{L}$ as below:

$$\mathcal{L} = L_{\text{comb}}(\hat{D}, D^{gt}) + \mu L_{\text{SI}}(\hat{D}_I, D^{gt}). \quad (8)$$

where μ is a balance term and empirically set to 0.1.

Training Details. We utilize a SoTA monocular depth estimation method, termed DepthFormer [34], as a primary backbone to validate our method effectively transfer the knowledge of large-scale depth model into our sensor-agnostic model. Our framework is implemented in public PyTorch [47], trained for 25 epochs on four RTX 3090TI GPUs using Adam [28] optimizer, with 228 × 304 and 240 × 1216 input resolution of NYU and KITTI dataset, respectively. Note that we resize the input RGB images to keep their ratio of height and width toward the foundation model used. The initial learning rate is 2×10^{-3}, and then scaled down with coefficients 0.5, 0.1, and 0.05 every 5 epochs after 10th epoch. The total training process for the NYU dataset takes approximately half a day, with an inference time of 0.06 seconds. For the KITTI dataset, the training time is about 1.5 days, with an inference time of 0.38 seconds. The framework comprises 53.4 million learnable parameters, which includes 0.1M dedicated to tuning the foundational model.

4. Experiment

In this section, we conduct comprehensive experiments to evaluate the impact of our depth prompting module on sensor-agnostic depth estimation. First, we briefly describe the experimental setup (Sec. 4.1), and present comparative results against various state-of-the-art (SoTA) methods on standard benchmark datasets (Sec. 4.2). Moreover, we provide an in-depth examination of bias issues in sensor (Sec. 4.3) as well as an ablation study to demonstrate the effect of each component in our method (Sec. 4.4). Lastly, we offer qualitative results to show zero generalization of our method on commercial sensors (Sec. 4.5).

4.1. Experiment Setup

Evaluation Protocols. For our comparative experiments, we select a range of SoTA methods for depth estimation from sparse measurements. These include a series of spatial propagation networks such as CSPN [9], S2D [41], NLSPN [45], DySPN [35], CostDCNet [26], and CompletionFormer [73]. Additionally, we choose SAN [19], which are designed to adapt various sparse setups. We use common quantitative measures of depth quality: root mean square error (RMSE, unit: meter), mean absolute error (MAE, unit: meter), and inlier ratio (DELTA1, $\delta < 1.25$).

Datsets: NYUv2 and KITTI DC. We utilize the NYU Depth V2 dataset, an indoor collection featuring 464 scenes captured with a Kinect sensor. Following the official train/test split, we use 249 scenes (about 50K samples) in training phase, and 215 scenes (654 samples) are tested for the evaluation. The NYU Depth V2 dataset provides 320×240 resolutions. We use the center-cropped image whose resolution is 304×228 and randomly sample 500 points to simulate the sparse depth.

For outdoor scenarios, we choose a KITTI DC [59] dataset with 90K samples. Each sample includes color images and aligned sparse depth data (about 5% density over image resolution) captured using a high-end Velodyne HDL-64E LiDAR sensor. The images have 1216×352 resolution. The dataset is divided into training (86K samples), validation (7K samples), and testing segments (1K samples). Ground truth is established by accumulating multiple LiDAR frames and filtering out errors, which results in denser LiDAR depths (about 20% density).

4.2. Experimental Results

Sensor Agnsoticity. We assess the versatility of our method and the SoTA methods across various density levels. They are commonly trained with a standard training protocol, e.g., 500 random depth samples from the NYUv2 dataset and 64 lines on the KITTI DC dataset. We test them under exactly the same conditions. For the NYUv2 dataset, we sample fewer samples (from 200 to 1 depth point) than that used in training phase. In addition, we use less scanning lines (from 32 to 1 line) than the original KITTI dataset.

As shown in Tabs. 1 and 2, our method consistently provides the superior results in almost test conditions. While methods such as NLSPN [45] and CompletionFormer [73] demonstrate their robustness with 200 samples in NYUv2 dataset and the 32-line scenario in KITTI dataset, respectively, our approach outperform the SoTA models in the

# Samples	200			100			32			8			4			1		
	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1
CSPN	0.1563	0.0707	0.9868	0.2795	0.1491	0.9585	0.6306	0.4344	0.7310	0.9688	0.7417	0.5031	1.0399	0.8186	0.4473	1.1093	0.8850	0.4108
S2D	0.1871	0.1031	0.9829	0.2733	0.1611	0.9598	0.4084	0.2615	0.9025	1.0982	0.8236	0.4537	1.7385	1.4294	0.1787	1.8446	1.5477	0.1416
NLSPN†	0.1358	**0.0553**	**0.9899**	0.2452	0.1125	0.9693	0.5541	0.3427	0.8254	0.9564	0.7023	0.5479	1.0775	0.8273	0.4511	1.1929	0.9521	0.3616
DySPN	0.1532	0.0686	0.9880	0.3174	0.1799	0.9345	0.6603	0.4838	0.6751	0.9635	0.7586	0.4913	1.0351	0.8304	0.4484	1.1079	0.9014	0.4110
CostDCNet†	0.1455	0.0606	0.9887	0.2809	0.1300	0.9592	0.6887	0.4144	0.7735	1.1685	0.8479	0.4864	1.3097	0.9973	0.3915	1.4255	1.1297	0.3065
CompletionFormer†	**0.1352**	0.0583	0.9898	0.3553	0.2125	0.9069	0.7921	0.6160	0.5250	1.0647	0.8536	0.3830	1.1259	0.9112	0.3526	1.2091	0.9878	0.3167
Ours	0.1435	0.0642	0.9881	**0.1778**	**0.0870**	**0.9812**	**0.2472**	**0.1452**	**0.9546**	**0.3673**	**0.2464**	**0.9133**	**0.3827**	**0.2627**	**0.9031**	**0.4040**	**0.2848**	**0.8935**

Table 1. Quantitative Results on NYUv2. † indicates that the publicly available code and model weight is utilized in this experiment.

# Lines	32			16			8			4			2			1		
	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1
CSPN	1.3644	0.4008	0.9876	2.0935	0.7532	0.9596	3.5806	1.6679	0.8580	5.7438	3.1650	0.6656	8.9143	5.6544	0.4478	12.5666	8.2679	0.3379
S2D	1.8133	0.7426	0.9623	2.6886	1.1670	0.9232	4.5806	2.6697	0.6727	7.2951	4.9223	0.3866	10.1624	6.8438	0.2854	12.2972	7.9316	0.2057
NLSPN†	1.1894	0.3536	0.9923	1.9279	0.6976	0.9675	3.2285	1.5482	0.8692	4.7571	2.5976	0.7267	6.0305	3.8779	0.4904	8.8244	5.2859	0.3949
DySPN	1.6758	0.5449	0.9871	2.3979	0.9096	0.9624	3.4687	1.5774	0.883	5.2374	2.8549	0.6981	6.5413	4.0182	0.5118	9.5199	5.3260	0.4637
CompletionFormer†	1.2513	0.3844	0.9912	2.1857	0.8403	0.9627	3.6505	1.7687	0.8577	6.2532	3.4800	0.6787	8.9682	5.9899	0.4672	12.7693	9.0019	0.3414
SAN†	1.8188	0.8160	0.9793	2.8866	1.5915	0.9087	3.7936	1.8339	0.8967	4.5894	2.3092	0.8523	4.1416	2.1280	0.8591	4.4444	2.3270	0.8153
Ours	**1.1465**	**0.3472**	**0.9942**	**1.3512**	**0.4134**	**0.9922**	**1.6419**	**0.5470**	**0.9888**	**1.9507**	**0.7629**	**0.9809**	**2.3841**	**1.1976**	**0.9505**	**2.8234**	**1.2678**	**0.9535**

Table 2. Quantitative Results on KITTI DC. † indicates that the publicly available code and weight are used in this experiment.

KITTI ↓ NYU	100-shot				10-shot			
	500		50		500		50	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CSPN	0.1848	0.0909	1.0235	0.7220	0.6741	0.5038	1.2859	1.0251
NLSPN	0.1940	0.1075	0.5917	0.4285	0.7367	0.5682	1.2798	1.0378
CompletionFormer	0.2259	0.1287	1.1560	0.4589	0.5881	0.8758	1.3870	1.1430
Ours	**0.1748**	**0.0796**	**0.4697**	**0.3048**	**0.5466**	**0.4284**	**0.7990**	**0.6482**
NYU ↓ KITTI	100-shot				10-shot			
	64		8		64		8	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CSPN	1.2803	0.3645	5.8021	3.4752	2.2717	0.9287	9.9108	6.1467
NLSPN	1.5156	0.5194	4.5913	2.7022	3.6842	1.8959	10.8715	5.7224
CompletionFormer	1.3404	0.3770	5.0787	3.2564	2.3280	1.0014	10.8328	7.0782
Ours	**1.2752**	**0.3534**	**4.5874**	**2.4834**	**2.1108**	**0.7846**	**5.6511**	**3.1699**

Table 3. Cross-validation between Indoor and Outdoor dataset.

more challenging scenarios. The depth prompt encoder contributes to constructing an adaptive representation for randomly given seeds, regardless of the pattern and density.

We observe that a majority of SoTA methods heavily depend on predefined input configurations. Spatial propagation, a prevalent technique among these methods, relies on relations among neighboring pixels, requiring a substantial number of seeds to cover an entire scene. This dependency results in significant performance deterioration in scenarios with sparser initial depth seeds. In addition, SAN [19], which are engineered to merge depth and image features at the late fusion to achieve stability in varying sparsity conditions, also encounter the performance drop in Tabs. 1 and 2.

In contrast, thanks to the knowledge of the foundation model and the depth-oriented prompt engineering, our method achieves relatively stable performance. Our prompt module enables the construction of an adaptive affinity map according to the distribution of input data, whose effectiveness is enlarged by the zero-shot generalization for unseen visual attributes and data distributions.

Cross-validation between Indoor and Outdoor. To conduct a cross-validation between outdoor and indoor scenarios, we finetune our model and the comparison methods with only 10 and 100 images. Since active sensors provide metric depth to the model, the domain adaption via a few ground-truth level annotations is inevitable. Following [63], we randomly select the pair of images and depth data. As shown in Tab. 3, our method shows the superior performance than the SoTA methods, which demonstrates the model's effectiveness in preserving visual features across varying scan ranges and its successful adaptation of the pre-trained knowledge to different domains.

4.3. Sparsity, Pattern and Range Biases

To assess the effectiveness of our model against the sensor bias issues, we design experiments with varying conditions: sparsity (from 500 to 50 samples), patterns (from random to grid), and range changes (from 0m∼3m to 3m∼10m). We also design experiments for the outdoor scenario that vary across three key conditions: sparsity (from 64-Line to 8-Line), patterns (from line to random), and range changes (from 0m∼15m to 15m∼80m). For a fair comparison, all models do not conduct RDA for this experiment, including our proposed method.

Tabs. 4 and 5 reveals that the previous methods face significant challenges on the bias issues. In addition, as demonstrated in Tabs. 1 and 2, the density changes also lead to significant performance drop, particularly the dense to sparse scenario. In contrast, to address the *sparsity* bias, our prompt-based method constructs adaptive relations among pixels to properly propagate even in the changing conditions.

Next, we investigate the negative impact of *pattern* bias. We observe that a model trained on depth data with a certain pattern suffers from limited generality due to the incompatibility with abundant representation learned for latent spaces of other depth patterns. As shown in Tab. 4, we attribute this to positional information combined with image features. The comparison models, being continuously exposed to a fixed

	Sparsity (500→50)			Sparsity Rev. (50→500)			Pattern (Random→Grid)			Pattern Rev. (Grid→Random)			Range (3m∼10m→0m∼3m)			Range Rev. (0m∼3m→3m∼10m)		
	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1
CSPN	0.4902	0.3102	0.8323	0.1538	0.0802	0.9877	0.1108	0.0468	0.9937	0.7657	0.5401	0.6310	0.3419	0.2235	0.8055	0.2869	0.1533	0.9466
S2D	1.3680	1.0730	0.3006	0.5608	0.3134	0.8827	0.4794	0.4331	0.6836	0.8988	0.6230	0.6265	0.8430	0.6048	0.6208	0.8322	0.5514	0.7678
NLSPN	0.4639	0.3018	0.8554	0.1516	0.0758	0.9882	0.1136	0.0466	0.9933	1.2200	0.9202	0.3982	0.3758	0.2553	0.7871	0.3753	0.2329	0.9207
DySPN	0.4473	0.2700	0.8832	0.1487	0.0779	**0.9887**	0.1088	0.0439	0.9935	0.6700	0.4117	0.7461	0.3908	0.2654	0.7693	0.3891	0.2138	0.9237
CostDCNet	0.4701	0.2946	0.8569	0.1458	0.0717	0.9883	0.1248	0.0557	0.9921	0.4164	0.2649	0.8955	0.2160	0.1265	0.9449	0.2205	0.0992	**0.9788**
CompletionFormer	0.4776	0.2957	0.8510	0.1486	0.0754	0.9879	0.1183	0.0476	0.9925	0.8862	0.5993	0.6276	0.3486	0.2347	0.8207	0.6187	0.3713	0.8614
Ours+MiDaS	0.4472	0.2787	0.8567	0.1722	0.0754	0.9827	0.1403	0.0549	0.9884	0.4478	0.2840	0.8608	0.2334	0.1257	0.9691	0.2803	0.1252	0.9574
Ours+KBR	**0.3632**	**0.2282**	**0.8939**	0.1503	0.0651	0.9865	0.1170	0.0449	0.9922	0.3133	0.1980	0.9434	**0.2007**	**0.1024**	0.9697	**0.2110**	**0.0945**	0.9776
Ours	0.3997	0.2418	0.8825	**0.1453**	**0.0634**	0.9874	**0.1081**	**0.0419**	**0.9937**	**0.2961**	**0.1766**	**0.9291**	0.2060	0.1075	**0.9701**	0.2328	0.0958	0.9693

Table 4. Case study on *sparsity*, *pattern*, and *range* biases on the NYUv2 dataset.

	Sparsity (64→8)			Sparsity Rev. (8→64)			Pattern (Line→Random)			Pattern Rev. (Random→Line)			Range (15m∼80m→0m∼15m)			Range Rev. (0m∼15m→15m∼80m)		
	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1	RMSE	MAE	DELTA1
CSPN	3.943	1.986	0.827	1.256	0.373	0.989	0.798	0.304	0.998	1.635	0.380	0.985	11.283	4.978	0.726	8.949	6.751	0.415
S2D	4.814	2.918	0.682	2.721	1.352	0.963	2.007	1.467	0.835	3.222	1.788	0.921	11.438	6.499	0.406	11.418	8.918	0.300
NLSPN	5.052	2.771	0.774	1.539	0.718	0.986	1.116	0.553	0.995	1.889	0.412	0.983	11.857	7.580	0.204	13.727	11.186	0.221
DySPN	4.106	2.136	0.786	1.269	0.394	0.991	0.837	0.317	0.997	1.856	0.489	0.982	11.127	5.028	0.700	9.969	7.641	0.379
CompletionFormer	4.180	2.175	0.820	1.963	0.942	0.966	0.867	0.402	0.997	1.689	0.387	0.984	11.492	5.422	0.626	9.969	7.641	0.379
Ours+MiDaS	4.463	2.256	0.797	1.182	0.337	0.991	0.836	0.318	0.996	1.637	0.372	0.985	13.272	5.868	0.714	5.763	3.935	0.479
Ours+KBR	3.780	1.719	0.874	1.180	0.339	0.993	0.944	0.257	0.995	1.689	0.381	0.985	12.032	5.479	0.705	5.790	4.291	0.483
Ours	**2.453**	**1.047**	**0.967**	**1.037**	**0.302**	**0.994**	**0.704**	**0.206**	**0.998**	**1.617**	**0.361**	**0.986**	**7.648**	**3.354**	**0.759**	**1.994**	**1.072**	**0.890**

Table 5. Case study of *sparsity*, *pattern*, and *range* biases on the KITTI DC dataset.

pattern like a grid shape, are limited to its generalization. When image and depth information are jointly represented, this issue is further exacerbated. On the other hand, from the results, we find out that random sampling offers benefits akin to augmentation effects, making the model generality better. Our method allows the transfer of depth information trained from various sparse patterns to the model, which provides the same effect as random sampling.

Lastly, we check the *range* bias. In the training phase, we only use depth data whose maximum depth range is 3m. All the models are tested using depth data whose min/max range of the depth distribution is set to [3m, 10m]. As shown in Tab. 4, it becomes evident that most methods exhibit poor generalization performance. Notably, the CompletionFormer [73] and NLSPN [45] struggle to produce the general performance for the near and far regions. Our framework effectively tackles the challenge using the foundational model designed for monocular depth estimation. Based on the foundation model, which predicts relative depth maps for all pixels, our method infers absolute depth maps, which extends the sensor's limited scan ranges.

Fig. 4 shows a significant distinction between our method and others in depth map reconstruction. While the comparison methods face challenges in accurately representing scene depth, especially in areas where input seeds are provided, our method excels in reconstructing the entire depth map. One notable observation is about the scenarios involving the changes in scanning ranges (Fig. 4-(c)). Our method uniquely overcomes the common bias problem. This better performance is attributed to the strengths of our foundation model's knowledge and the sensor-adaptive depth prompt.

	Sparsity	Pattern	Range	Param.	Inference
w/o SPN Eq.(1)	0.498 / 0.334	0.145 / 0.096	0.546 / 0.379	53.3M	61.7ms
w/o Prompt Eq.(2)	0.452 / 0.288	0.301 / 0.207	0.686 / 0.551	49.7M	54.9ms
w/o Pretrain Eq.(3)	0.409 / 0.249	0.118 / 0.049	1.283 / 0.934	326.9M	64.4ms
w/o LS-solver Eq.(4)	0.416 / 0.268	0.118 / 0.052	0.520 / 0.305	53.4M	61.3ms
w/ RDA	**0.231 / 0.134**	0.113 / 0.046	0.426 / 0.251	53.4M	63.9ms
Ours	0.400 / 0.242	**0.108 / 0.0420**	**0.206 / 0.108**	53.4M	63.9ms

Table 6. Ablation study of our proposed methods (RMSE / MAE).

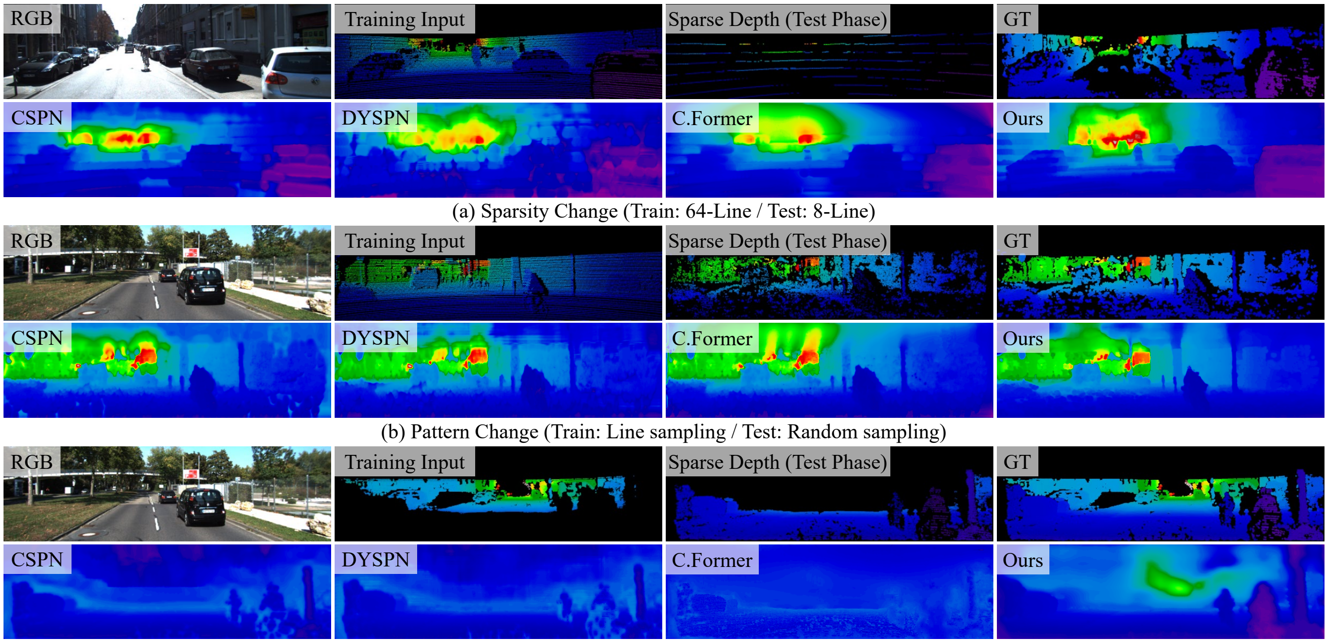
	NYU 100	NYU 8	NYU 1	KITTI 16	KITTI 4	KITTI 1
NLSPN	**0.178** / 0.089	0.434 / 0.290	0.649 / 0.491	1.662 / 0.620	2.307 / 0.930	3.271 / 1.464
C.Former	0.182 / 0.090	0.434 / 0.289	0.648 / 0.487	2.179 / 0.796	3.291 / 1.363	5.428 / 2.457
Ours	**0.178 / 0.087**	**0.367 / 0.246**	**0.404 / 0.285**	**1.351 / 0.413**	**1.951 / 0.763**	**2.823 / 1.268**

Table 7. Adaption RDA method to other methods (RMSE / MAE).

4.4. Ablation Study

Adaption to Foundation Models. To evaluate the versatility of our method with various foundational models, we replace our primary backbone with MiDaS [4] and KBR [56]. The MiDaS and KBR are developed for relative scale depth using large-scale datasets as well. Tab. 4 reveals that KBR outperforms other methods, including our own variant using the MiDaS backbone. We argue that the self-supervised training of KBR, unlike the supervised manner of MiDaS, provides a more generalizable feature space, dealing with challenging conditions [8, 61].

Component Ablation of the Proposed Method. We perform an additional ablation study on each component of our model in Tab. 6. The study reveals that the RDA method notably reduces the sparsity bias (w/ RDA). For the range bias, the pre-trained knowledge from the backbone contributes to performance improvement (w/o Pretrain). Our depth-oriented prompt engineering contributes to the overall performance (w/o Prompt). Additionally, the results of LS



(a) Sparsity Change (Train: 64-Line / Test: 8-Line)

(b) Pattern Change (Train: Line sampling / Test: Random sampling)

(c) Range Change (Train: 15m~80m / Test: 0m~15m)

Figure 4. Qualitative results for the changes of measurement patterns, sparsity, and scan ranges. We visualize images, input examples in the training phase, sparse depths in the test phase, and GT in the first row.

solver Eq.(4) show that the sensor biases are not addressed via the naive scale fitting, but are solved by the combinations of our components. As a solution, our idea is to exploit the SPN module to use the initial dense depth, which is aligned relative depth from the backbone with sparse absolute-scale depth (w/o SPN).

Random Depth Augmentation. RDA is an effective strategy to mitigate the issue of sparsity bias. To evaluate its compatibility and adaptability with other methods, we conduct an ablation study incorporating RDA into NLSPN [45], and CompletionFormer [73]. The results, as described in Tab. 7, demonstrate notable performance improvements in sparse input scenarios. This highlights that the RDA not only naturally improves the models' ability to generalize across different levels of data sparsity, but also becomes more effective when used together with prompt engineering.

4.5. Zero-shot Inference on Commercial Sensors

We verify our method's zero-shot generality by testing it on different datasets without any additional training. We use our model trained on NYUv2 [55] and KITTI DC [59] datasets, then apply it to dataset taken from various sensors such as Apple LiDAR [40], Intel RealSense [27], and 32-Line Velodyne LiDAR [54]. Here, for the Apple LiDAR dataset, we directly capture a set of indoor images using iPad Pro. As shown in Fig. 1, our method is applicable for Apple LiDAR compared to ARKit [40], which is a built-in framework on iOS devices. Additionally, it shows better generality with consistent results in the VOID dataset [64], collected using a stereo sensor in RealSense. Remarkably, our model, initially trained on 64-Line Velodyne LiDAR, excels in handling the NuScenes dataset [64] captured with fewer channel LiDAR over the second best approach in Sec. 4.2. Note that we describe details about the experimental setup, more quantitative and qualitative results, and further analysis in the supplemental materials.

5. Conclusion

We introduce a novel depth prompting technique, leveraging large-scale pre-trained models for high-fidelity depth estimation in metric scale. This approach significantly addresses the challenges of well-known sensor biases associated with fixed sensor densities, patterns, and limitation of range and enables sensor-agnostic depth prediction. Through the comprehensive experiments, we demonstrate the stability and generality of our proposed method, showcasing its superiority over existing methodologies. Furthermore, we verify our method on a variety of real-world scenarios through zero/few-shot inference across diverse sensor types.

Acknowledgement This research was supported by 'Project for Science and Technology Opens the Future of the Region' program through the INNOPOLIS FOUNDATION funded by Ministry of Science and ICT (Project Number: 2022-DD-UP-0312), GIST-MIT Research Collaboration grant funded by the GIST in 2024, the Ministry of Trade, Industry and Energy (MOTIE) and Korea Institute for Advancement of Technology (KIAT) through the International Cooperative R&D program in part (P0019797) and the Korea Agency for Infrastructure Technology Advancement(KAIA) grant funded by the Ministry of Land Infrastructure and Transport (Grant RS-2023-00256888).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2022. 2

[2] 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 (AAAI)*, 2023. 3

[3] Ayush Bhandari, Achuta Kadambi, Refael Whyte, Christopher Barsi, Micha Feigin, Adrian Dorrington, and Ramesh Raskar. Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *Optics letters*, 39(6):1705–1708, 2014. 2

[4] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 2, 3, 4, 7

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[7] Han Cai, Chuang Gan, Ligeng Zhu, and Song Han. Tinytl: Reduce memory, not parameters for efficient on-device learning. 2020. 2, 4

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 7

[9] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3, 5

[10] Andrea Conti, Matteo Poggi, and Stefano Mattoccia. Sparsity agnostic depth completion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5871–5880, 2023. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] Eric Dexheimer and Andrew J Davison. Learning a depth covariance function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[13] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 2, 3

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. 2014. 5

[16] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4), 2022. 3

[17] 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, 4

[18] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023. 2

[19] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 5, 6

[20] Xiaoyang Guo, Hongsheng Li, Shuai Yi, Jimmy Ren, and Xiaogang Wang. Learning monocular depth by distilling cross-domain stereo networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[22] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 4

[23] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, 2011. 1

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[25] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual

prompt tuning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 4

[26] Jaewon Kam, Jungeon Kim, Soongjin Kim, Jaesik Park, and Seungyong Lee. Costdcnet: Cost volume based depth completion for a single rgb-d image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 5

[27] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2017. 1, 8

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[30] Robert Lange, Peter Seitz, Alice Biber, and Stefan C Lauxtermann. Demodulation pixels in ccd and cmos technologies for time-of-flight ranging. In *Sensors and camera systems for scientific, industrial, and digital photography applications*, pages 177–188. SPIE, 2000. 1

[31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[32] Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019. 2

[33] Yuenan Li, Jin Wu, and Zetao Shi. Lightweight neural network for enhancing imaging performance of under-display camera. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 2023. 3

[34] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, pages 1–18, 2023. 2, 4, 5

[35] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 1, 2, 3, 5

[36] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 3

[37] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[38] Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, and Chao Dong. Degae: A new pretraining paradigm for low-level vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[39] Kaiyue Lu, Nick Barnes, Saeed Anwar, and Liang Zheng. From depth what can you see? depth completion via auxiliary image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[40] Gregor Luetzenburg, Aart Kroon, and Anders A Bjørk. Evaluation of the apple iphone 12 pro lidar for an application in geosciences. *Scientific reports*, 11(1):22221, 2021. 1, 8

[41] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018. 5

[42] Donald W Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963. 4

[43] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[44] Ayushya Pare, Shufang Zhang, and Zhichun Lei. Multipath interference suppression in time-of-flight sensors by exploiting the amplitude envelope of the transmission signal. *IEEE Access*, 8:167527–167536, 2020. 2

[45] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 5, 7, 8

[46] Jin-Hwi Park, Jaesung Choe, Inhwan Bae, and Hae-Gon Jeon. Learning affinity with hyperbolic representation for spatial propagation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 1, 2, 3

[47] 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. In *Proceedings of the Neural Information Processing Systems Workshop (NeurIPS-W)*, 2017. 5

[48] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[49] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. In *preprint*. OpenAI, 2018. 2

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 2

[52] Thomas Richter, Jürgen Seiler, Wolfgang Schnurrer, and André Kaup. Robust super-resolution for mixed-resolution multiview image plus depth data. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, 26(5): 814–828, 2015. 3

[53] Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135, 2021. 2

[54] Brent Schwarz. Mapping the world in 3d. *Nature Photonics*, 4(7):429–430, 2010. 1, 8

[55] 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. 2, 4, 8

[56] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Kick back & relax: Learning to reconstruct the world by watching slowtv. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 7

[57] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2020. 2

[58] Tommi Tykkälä, Cédric Audras, and Andrew I Comport. Direct iterative closest point for real-time visual odometry. In *Proceedings of the IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2011. 2

[59] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 1, 5, 8

[60] Attila Vidács and Géza Szabó. Winning ariac 2020 by kissing the bear: Keeping things simple in best effort agile robotics. *Robotics and Computer-Integrated Manufacturing*, 71:102166, 2021. 2

[61] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 7

[62] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[63] Yao Wei, Yanchao Sun, Ruijie Zheng, Sai Vemprala, Rogerio Bonatti, Shuhang Chen, Ratnesh Madaan, Zhongjie Ba, Ashish Kapoor, and Shuang Ma. Is imitation all you need? generalized decision-making with dual-phase training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6

[64] 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. 8

[65] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[66] Lu Xia, Chia-Chih Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2012. 1

[67] Zhihao Xia, Patrick Sullivan, and Ayan Chakrabarti. Generating and exploiting probabilistic monocular depth estimates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[68] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, and Chunhua Shen. Towards domain-agnostic depth completion. *arXiv preprint arXiv:2207.14466*, 2022. 2, 3

[69] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[70] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(10):6360–6376, 2022. 3

[71] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[72] Song Zhang. High-speed 3d shape measurement with structured light methods: A review. *Optics and lasers in engineering*, 106:119–131, 2018. 1

[73] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 7, 8

[74] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1

[75] Yichao Zhou, Haozhi Qi, Yuexiang Zhai, Qi Sun, Zhili Chen, Li-Yi Wei, and Yi Ma. Learning to reconstruct 3d manhattan wireframes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1