Dense Metric Depth Completion from Sparse Direct Time-of-Flight Sensors

Hakyeong Kim*
KAIST
hkkim@vclab.kaist.ac.kr

Ruicheng Wang*
USTC
t-ruiwang@microsoft.com

Chengtang Yao
Microsoft Research Asia
chengtangyao@microsoft.com

Jiaolong Yang
Microsoft Research Asia
jiaoyan@microsoft.com

Min H. Kim
KAIST
minhkim@vclab.kaist.ac.kr

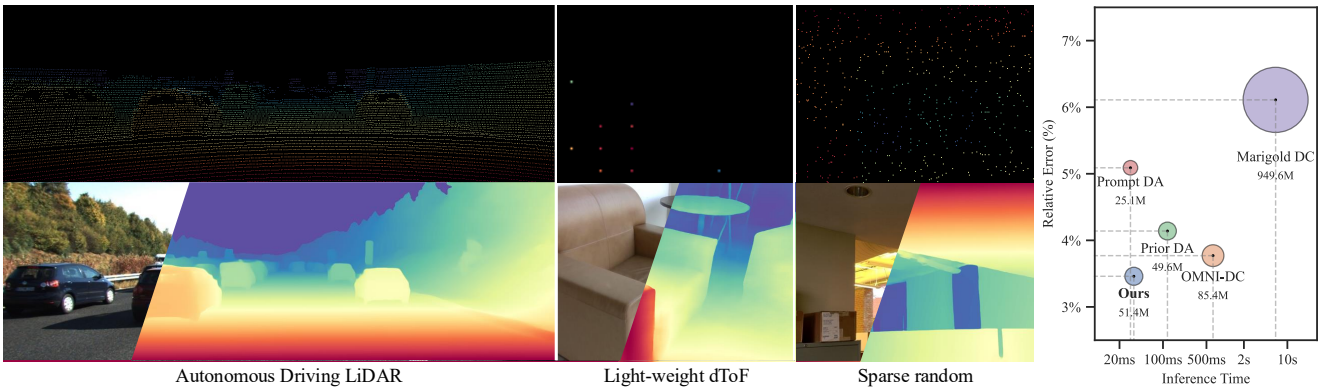

Figure 1. Zero-shot generalization of our model across different dToF sensing conditions. Top: sparse depth inputs from three representative settings—autonomous driving LiDAR (rotating dToF), lightweight mobile dToF, and extremely sparse random sampling. Bottom: our predicted dense metric depth for each case, demonstrating strong robustness under diverse sparsity, noise, and sensor patterns. Right: comparison of accuracy versus inference time (bubble size indicates model parameters). Our method achieves the most favorable balance of low error and fast runtime, outperforming state-of-the-art depth completion and enhancement approaches. See Table 1 for numerical details.

Abstract

Direct Time-of-Flight (dToF) sensors provide highly accurate metric depth and are more robust than indirect ToF systems in challenging real-world conditions. However, their high manufacturing cost and limited photodiode array size produce depth maps that are extremely sparse, low-resolution, and noisy, making them unsuitable for VR/XR, robotics, and 3D perception tasks that require dense metric depth. Existing monocular and depth completion methods struggle to handle the unique sampling patterns and hardware artifacts of dToF devices, and their performance often deteriorates significantly under severe sparsity or noise. We present a generalizable framework for dense metric depth completion from sparse dToF measurements, capable of operating across diverse sensor types, sparsity levels, and noise conditions. Our model employs a depth-guided dual-branch Vision Transformer encoder that processes RGB images and sparse dToF measurements separately, while a masked joint attention module allows depth tokens to reliably guide image features without being overwritten by them. A lightweight decoder reconstructs dense metric depth efficiently, without diffusion-based or refinement-heavy post-processing. To address the scarcity of paired training data, we introduce a comprehensive dToF simulation pipeline that reproduces the characteristics of flash, sub-VGA flash, and rotating sensors, including hardware-induced degradation, irregular sparsity, and realistic noise distributions. Trained entirely on synthetic data, our model achieves strong zero-shot generalization across 6 datasets and 3 real dToF devices, outperforming state-of-the-art approaches in both accuracy and computational efficiency. This establishes a robust and practical solution for dense metric depth completion from sparse direct ToF sensors. Our code and models will be open-sourced. See project page.

1. Introduction

Dense and high-quality depth perception is essential for VR/XR systems, robotics, and general 3D scene understanding. Although recent advancements in monocular depth esti-

*Work done during internship at Microsoft Research Asia.

mation have shown impressive in-the-wild performance by training large-scale foundation models [13, 15, 37, 38, 44], these approaches still suffer from scale ambiguity and often produce inconsistent metric depth in complex real-world scenarios. This limitation makes them unreliable for applications that require precise and stable metric depth.

A common strategy to mitigate scale ambiguity is to incorporate sparse direct Time-of-Flight (dToF) measurements, such as LiDAR or low-resolution ToF depth, which provide accurate range information. While this integration has significantly improved robustness in various real-world settings [19, 21, 24], existing methods are typically tailored to a specific depth sensor and struggle when the measurements become extremely sparse, noisy, or irregularly sampled. Furthermore, recent attempts that rely on diffusion models [34], iterative optimization [50], or multi-stage refinement [41] achieve strong performance but incur substantial computational overhead, limiting their practicality in mobile or real-time applications.

In this paper, we introduce a single model that generalizes effectively to diverse sparse dToF devices and challenging real-world scenarios, even under extreme sparsity or high noise, while remaining computationally efficient. We observe that many prior methods rely heavily on pretrained monocular encoders and treat sparse depth merely as an auxiliary signal, failing to capture the complementary and mutually constraining structure shared between RGB features and dToF measurements. To address this, we propose a depth-guided dual-branch encoder that independently processes RGB images and sparse dToF depth, while a masked joint attention module enables controlled information flow from depth tokens to RGB tokens without corrupting the depth representation. This design yields a more expressive and depth-aware feature space, allowing a lightweight decoder to produce high-quality dense metric depth without requiring complex refinement networks.

To further enhance generalization, we introduce a comprehensive dToF simulation pipeline that models a wide range of sensor characteristics, including flash, and rotating devices, along with realistic noise patterns, occlusion-induced missing regions, and hardware-level degradation. This scales training to large synthetic datasets and learns a robust representation that transfers well to real-world sparse dToF inputs.

We conduct extensive experiments across 6 datasets, 3 dToF sensor devices, and multiple sparsity and noise levels, covering indoor scenes and street environments. Our pretrained model exhibits strong zero-shot generalization across all settings and outperforms state-of-the-art depth completion and fusion-based approaches in both accuracy and efficiency. Ablation studies further validate the effectiveness of our encoder architecture.

Our contributions are summarized as follows:

- We introduce a new framework for dense metric depth completion from sparse direct ToF measurements, achieving strong zero-shot generalization across diverse sensor devices and depth-related tasks using a single model trained solely on synthetic data.
- We propose a depth-guided dual-branch encoder with masked joint attention, enabling effective and controlled feature exchange between RGB and sparse depth. This expressive encoder allows us to use a lightweight decoder without complex refinement while maintaining high accuracy and efficiency.
- We design a comprehensive dToF simulation pipeline that reproduces realistic sensor noise, sparsity patterns, and hardware artifacts across multiple dToF device families, significantly improving synthetic-to-real transfer.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation aims to infer 3D scene geometry from a single RGB image. Recent advances have demonstrated strong generalization in the wild by scaling up depth foundation models [37, 43, 44] or by leveraging diffusion-based generative priors [6, 11, 15]. Although these approaches achieve high-quality relative depth, they are typically trained with scale-invariant losses and therefore cannot reliably recover absolute metric scale, which is required in many downstream applications.

To address scale ambiguity, several works estimate metric depth by direct scale regression from image features [2, 25, 38], while others perform scale alignment using camera intrinsics or geometric constraints [9, 13, 47]. Despite the more explicit treatment of scale, these methods still struggle with metric accuracy under challenging lighting, textureless regions, and complex outdoor conditions.

In contrast to monocular-only methods, we incorporate sparse dToF measurements as reliable geometric references. This allows our model to preserve absolute scale and produce high-quality dense metric depth across diverse real-world scenarios.

2.2. Depth Completion

Depth completion seeks to infer dense depth from sparse measurements. Early methods densify sparse inputs by propagating depth values using hand-crafted priors or convolutional networks [3, 12, 20, 30, 40]. While effective, these models are usually designed for specific sensors or datasets, limiting generalization when applied to new environments or different sparsity patterns.

Recent studies aim for broader generalization by combining sparse depth with powerful monocular priors. PromptDA [19] and DepthPrompting [24] integrate monocular depth foundation models into completion pipelines, achieving strong zero-shot performance but remaining tied to specific

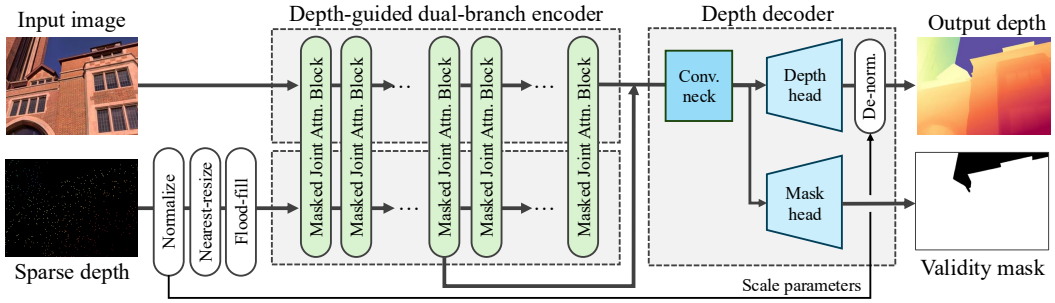

Figure 2. Overview of our method. Given an input RGB image and sparse dToF depth, we first normalize and upsample the sparse measurements to obtain a unified depth representation. The RGB image and sparse depth are then encoded by a dual-branch ViT encoder, where multiple masked joint attention blocks enable depth tokens to guide RGB features without corrupting the depth representation. The fused features are passed to a lightweight DPT decoder, which predicts dense normalized depth and a validity mask. The final metric depth is recovered through de-normalization using scale parameters from preprocessing. Our design produces high-quality dense metric depth from sparse dToF inputs while remaining computationally efficient.

sensor configurations. Diffusion-based approaches such as Marigold-DC [34] and DepthLab [21] further improve generalization without task-specific training, yet they are sensitive to noise and extremely sparse inputs, and often require heavy computation.

Optimization-based frameworks such as OGNI-DC [49] and OMNI-DC [50] perform iterative refinement at inference time, offering robustness but with significant runtime overhead. PriorDA [41] improves prediction quality by applying monocular priors in both coarse alignment and fine refinement stages, but this two-stage pipeline is computationally expensive.

Unlike the above methods, our approach does not rely on diffusion models, iterative optimization, or multi-stage refinement. Instead, we demonstrate that a highly efficient dual-branch Transformer encoder, combined with realistic dToF simulation, can achieve strong zero-shot generalization across diverse dToF devices, noise conditions, and sparsity levels while maintaining fast inference and practical applicability.

3. Method

Given a high-resolution RGB image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ and a sparse depth map $\boldsymbol{Z} \in \mathbb{R}^{H\times W}$ captured by a dToF sensor, our goal is to estimate a dense, high-resolution metric depth map $\tilde{\boldsymbol{D}} \in \mathbb{R}^{H\times W}$ along with a validity mask $\tilde{\boldsymbol{M}} \in \mathbb{R}^{H\times W}$ indicating unreliable or unobservable regions (e.g., sky), as illustrated in Figure 2. To accommodate diverse dToF sensor devices, we first preprocess the sparse depth input to construct a unified depth representation. We then employ a depth-guided dual-branch encoder that separately encodes RGB and sparse depth features while enabling controlled cross-modal fusion through a masked joint attention mechanism. This produces depth-aware image features that are subsequently decoded by a lightweight DPT-based decoder to generate dense metric depth. Lastly, we introduce a comprehensive synthetic data generation pipeline that models diverse dToF sensor characteristics, noise distributions, and sparsity patterns, significantly improving the model's robustness and zero-shot generalization.

3.1. Preprocessing

Depth projection and filling. To address the heterogeneous resolutions and sampling patterns of different dToF devices, we first convert the raw sensor depth into a unified representation. Because dToF depth maps typically have much lower spatial resolution than the corresponding RGB image, we upsample both the depth map and its validity mask to the RGB resolution. We then fill missing-depth regions using nearest-neighbor interpolation to obtain a continuous depth field that preserves local geometric structures. Importantly, the validity mask remains sparse, allowing the network to distinguish between pixels originating from actual sensor measurements and those generated through interpolation.

Depth normalization. To enable the depth branch to use the pretrained ViT encoder from DINOv2 [23], we convert the sparse depth map into a three-channel tensor whose statistical distribution matches that of DINOv2's RGB input. We first apply log-normalization to the sensor depth $\boldsymbol{Z}$ to stabilize scale variations and compress large depth ranges:

$$\begin{aligned} \alpha &= \log(Z_{\max}) - \log(Z_{\min}), \\ \beta &= \log(Z_{\min}), \\ \hat{\boldsymbol{Z}} &= (\log \boldsymbol{Z} - \beta)/\alpha, \end{aligned} \tag{1}$$

where $Z_{\min}$ and $Z_{\max}$ are the raw minimum and maximum valid depth measurements, respectively. In practice, this normalization remains stable under noisy and outlier-heavy depth simulations.

We then form a three-channel depth input by duplicating the normalized depth for the first two channels and using the sparse validity mask as the third channel. This provides the encoder with both geometric intensity information and measurement reliability. The final tensor is linearly scaled to the range $[-1, 1]$ to align with the dynamic range expected by the pretrained ViT.

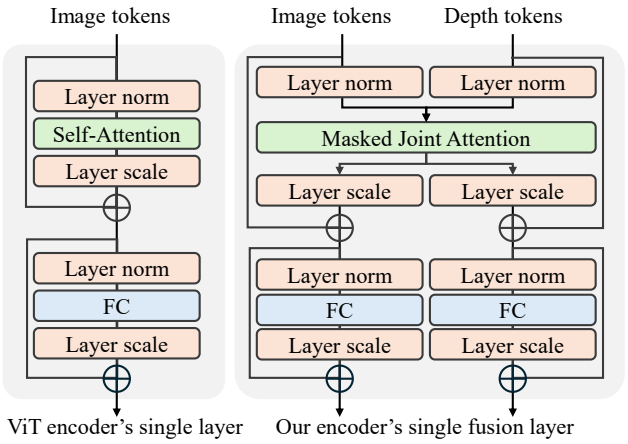

Figure 3. Left: a single layer of the standard ViT encoder, which performs self-attention and feed-forward operations on image tokens only. Right: our proposed fusion layer, where image and depth tokens are processed in two parallel branches and fused through a masked joint attention module. This design preserves the original ViT structure while enabling depth-guided feature interaction between the two modalities.

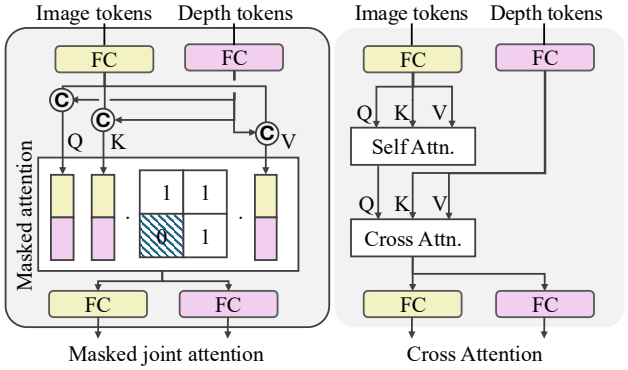

Figure 4. Left: in our masked joint attention, queries, keys, and values are computed jointly from the concatenated image and depth tokens. A directional mask is applied to restrict information flow from image tokens to depth tokens, while still allowing depth-to-image guidance. This enables effective and controlled fusion of spatial and geometric cues. Right: in cross-attention, queries are taken from one modality while keys and values come from the other, leading to asymmetric dependence and modality-specific attention flows.

3.2. Depth-Guided Dual-Branch Encoder

Most depth completion approaches rely on an off-the-shelf monocular encoder and treat sparse depth as an auxiliary input, overlooking the strong correspondence between RGB appearance and geometric cues. This often limits their robustness when the sparse depth becomes extremely irregular or noisy. In contrast, we design an encoder that explicitly models cross-modal interactions between RGB images and sparse dToF measurements.

Our encoder consists of two parallel Vision Transformer (ViT) branches [26], one for the RGB image and one for the normalized sparse depth input. Instead of processing the two modalities independently or relying on standard cross-attention, we adopt a masked joint attention mechanism that enables controlled information exchange between the branches (Figure 3).

In masked joint attention, tokens from the image $[Q_I, K_I, V_I]$ and depth branches $[Q_Z, K_Z, V_Z]$ are first concatenated, after which the queries, keys, and values are computed jointly:

$$Q = \begin{bmatrix} Q_I \\ Q_Z \end{bmatrix}, \quad K = \begin{bmatrix} K_I \\ K_Z \end{bmatrix}, \quad V = \begin{bmatrix} V_I \\ V_Z \end{bmatrix}.$$

This differs from conventional cross-attention, where queries and keys/values come from different modalities (Figure 4). We then apply a directional mask during the attention computation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top \odot G}{\sqrt{d_k}}\right) V, \; G = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad (2)$$

The mask G enforces an asymmetric information flow: depth-to-image attention is allowed, enabling depth measurements to guide and refine image features, while image-to-depth attention is suppressed, preventing unreliable RGB cues from corrupting the sparse depth representation. This yields depth-aware image embeddings while preserving the integrity of the depth features.

An additional advantage is that masked joint attention closely matches the structure of the original ViT self-attention block. As a result, the overall encoder remains equivalent to two isomorphic ViT branches, allowing us to initialize both with pretrained DINOv2 weights [23]. This initialization injects strong visual and geometric priors from large-scale pretraining, significantly improving both convergence and generalization.

3.3. Depth Decoder

Our depth decoder follows the DPT architecture [26, 37] and consists of two lightweight decoding branches. The first branch predicts a single-channel validity mask, while the second branch estimates a dense normalized depth map $\hat{\boldsymbol{D}}$. To recover metric depth, we apply the same de-normalization parameters α, β used during preprocessing (Eq. (1)):

$$\tilde{\boldsymbol{D}} = \exp(\alpha \cdot \hat{\boldsymbol{D}} + \beta). \quad (3)$$

Because $\hat{\boldsymbol{D}}$ is not constrained to the interval $[0, 1]$, this formulation allows the decoder to extrapolate depth values beyond the original sensor range $[Z_{\min}, Z_{\max}]$. As a result, the model can infer plausible metric depth even in regions where the dToF sensor provides no reliable measurements.

3.4. Loss

We supervise our model using four complementary loss terms: depth-weighted L1 loss $\mathcal{L}1$, global scale-invariant loss $\mathcal{L}g$, local scale-invariant loss $\mathcal{L}l$, and a mask loss $\mathcal{L}m$:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_g + \mathcal{L}_l + \mathcal{L}_m. \quad (4)$$

We use equal weighting for all loss terms, which is consistent with prior work's [37] design choice.

Depth-weighted L1 loss. We penalize the metric depth error between the predicted depth $\tilde{\boldsymbol{D}} = \{\tilde{d}_i\}$ and the ground truth $\boldsymbol{D} = \{d_i\}$ over all valid ground-truth pixels $\mathcal{M}$:

$$\mathcal{L}_1 = \sum_{i \in \mathcal{M}} \frac{1}{d_i} |\tilde{d}_i - d_i|_1, \tag{5}$$

The inverse-depth weighting prevents the model from overfitting to large far-range values and encourages accurate estimation in near-range regions where geometric structures are more detailed.

Global and local scale-invariant losses. While L1 loss ensures metric correctness, it is insensitive to geometric shape, which may lead to distortions in object structures. To preserve both global and local geometry, we compute scale-invariant losses in 3D space.

We first project the predicted and ground truth depth maps into 3D point clouds, $\tilde{\boldsymbol{P}} = \{\tilde{\boldsymbol{p}}_i\}$ and $\boldsymbol{P} = \{\boldsymbol{p_i}\}$, using camera intrinsics. We then estimate a global scale factor s using the ROE solver [37]. For local geometry, we divide the image into patches $\{\mathcal{S}_j\}_{j \in \mathcal{H}}$, where $\mathcal{H}$ is the set of sampled anchor points, and compute patch-specific scales s_j. The corresponding losses are:

$$\mathcal{L}_g = \sum_{i \in \mathcal{M}} \frac{1}{d_i} \|s\tilde{\boldsymbol{p}}_i - \boldsymbol{p}_i\|_1, \quad \mathcal{L}_l = \sum_{j \in \mathcal{H}} \sum_{i \in \mathcal{S}_j} \frac{1}{d_i} \|s_j\tilde{\boldsymbol{p}}_i - \boldsymbol{p}_i\|_1. \tag{6}$$

Together, these losses encourage the model to maintain a consistent metric scale while preserving fine-grained geometric structure.

Mask loss. Following MoGe [37], we also predict a validity mask $\tilde{\boldsymbol{M}} = \{\tilde{m}_i\}$ to identify regions with undefined or infinite depth (e.g., sky, reflective areas). The mask head is trained using binary cross-entropy:

$$\mathcal{L}_m = -\sum_i \left[m_i \log(\tilde{m}_i) + (1 - m_i) \log(1 - \tilde{m}_i) \right], \tag{7}$$

where $\boldsymbol{M} = \{m_i\}$ denotes the ground truth mask. This term ensures reliable uncertainty handling in regions where the sensor cannot provide depth measurements.

3.5. Sparse dToF Depth Simulation

High-resolution ground-truth depth paired with real dToF sensor measurements is extremely scarce and limited in diversity. To overcome this constraint, we adopt a simple yet powerful strategy: we synthesize sparse dToF depth from a wide range of large-scale synthetic RGB-D datasets [4, 5, 7, 10, 14, 17, 18, 22, 26–28, 31, 35, 36, 39, 42, 48]. This simulation pipeline enables us to scale training to millions of diverse scenes, effectively compensating for the lack of real sensor data and significantly improving robustness and zero-shot generalization.

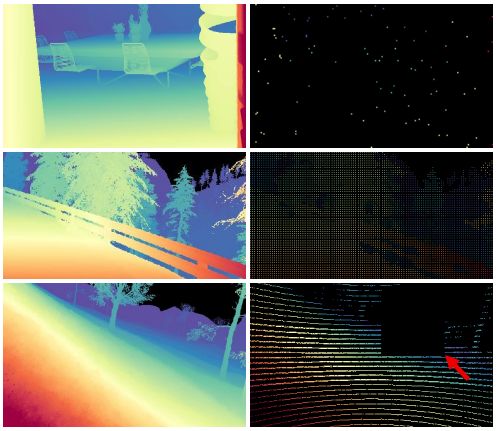

Figure 5. We simulate three representative patterns of direct ToF sensors: (top) extremely sparse Flash dToF, (middle) sub-VGA-resolution Flash dToF with image space projections, and (bottom) rotating LiDAR-style dToF with line-pattern sampling. The simulation reproduces realistic sparsity, resolution limits, and noise characteristics observed in real devices. The red arrow highlights an example of our depth inpainting augmentation, which models occlusions and missing returns encountered in real-world sensing.

Direct ToF sensors vary widely in their scanning mechanisms, which in turn determine the sparsity pattern, resolution, and noise characteristics of their depth outputs. We categorize these devices into two representative families: flash dToF, and rotating dToF. Each type exhibits distinct sampling distributions—from extremely sparse dot patterns to low-resolution dense grids and line-structured scans. To ensure compatibility with real devices, we simulate each family using sensor-specific rules that reproduce their unique sparsity, non-uniform coverage, and noise behavior (Figure 5). This diverse sensor-aware simulation plays a crucial role in teaching our model to handle the highly irregular and device-dependent patterns encountered in real-world dToF measurements.

Flash dToF camera. Flash dToF sensors, widely used in mobile devices, illuminate the entire scene at once and return depth values determined by the resolution of their SPAD array. This results in extremely sparse depth samples. To simulate this sensing pattern, we randomly sample 64–10K points from the ground-truth depth map, following real sensor specifications and recent depth completion settings [21, 34].

In addition to sparse flash sensors, many commercial devices output sub-VGA–resolution dense depth maps produced through onboard post-processing or by embedding small SPAD arrays in larger layouts [1, 45]. Although denser, these maps exhibit characteristic edge degradation, spatial smoothing, and hardware-induced artifacts. To model this second class of flash dToF sensors, we first downsample the ground-truth depth using nearest-neighbor interpolation to preserve thin structures. We then apply random erosion and dilation to depth boundaries to reproduce the blurry and

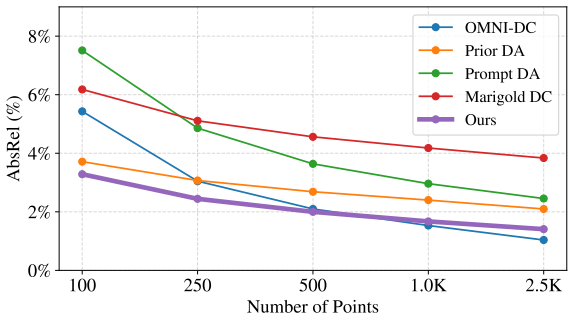

Figure 6. Quantitative analysis on varying number of depth inputs. Refer to supplements for numerical details.

softened edges commonly observed in real sensor output. To introduce spatially coherent variations, we generate Perlin noise at the image resolution and use it to assign a smoothly varying erosion–dilation radius to each pixel.

Rotating dToF camera. Rotating dToF sensors, widely used in automotive LiDAR systems, acquire depth by sweeping laser beams across the scene, producing line-structured and view-dependent depth patterns. To simulate this sensing mechanism, we follow the specifications of the Velodyne VLP-16 and VLP-32 families. For each simulated scan, we sample ray directions stochastically within the parameter ranges spanned by these devices, including horizontal resolutions of 0.33°–2°, vertical resolutions of 0.1°–0.4°, vertical fields of view from $-30°$ to 30°, and Gaussian noise levels up to a standard deviation of 0.001. This stochastic configuration yields realistic variations in scan density, angular spacing, and measurement noise, capturing the diverse sampling characteristics of real rotating dToF sensors.

Noise augmentations. In addition to modeling the characteristics of different dToF devices, we introduce a set of data augmentation strategies that simulate the degradation patterns commonly observed in real-world sensing conditions. dToF measurements frequently exhibit missing depth on highly reflective or low-return surfaces and under strong ambient illumination. Multipath interference around transparent or specular objects can produce erroneous depth values, while depth drift may occur due to object motion, sensor vibration, or imperfect calibration. Because these artifacts are highly device-dependent, we incorporate randomized perturbations to encourage our model to become device-agnostic.

Concretely, we add Gaussian noise proportional to depth magnitude, apply random spatial jitter, and inject 0.2–1.0% outlier points whose depths are uniformly sampled between the near and far planes of the ground-truth depth map. To further mimic occlusions and unobserved regions, we apply depth inpainting augmentation, which randomly removes square patches, edge-aligned regions, and irregular masks generated using Perlin noise. These augmentations collectively expose the model to a broad spectrum of real-world degradation patterns, significantly improving robustness to diverse sensing environments.

4. Experiments

4.1. Implementation Details

During preprocessing, we apply a flood-fill algorithm to complete missing-depth regions before constructing the unified depth representation. Our encoder is initialized with ViT-Small weights pretrained using DINOv2, and we extract features from the 6th and 12th intermediate layers to feed into the DPT-style decoder. Following MoGe-2 [38], the decoder adopts a multi-scale architecture, but we reduce the feature dimensions of the residual blocks to 384, 256, 64, 32, and 16 for improved efficiency. Our implementation also supports a dynamic number of depth tokens; during training, the number of RGB tokens is sampled from $[800, 2000]$, while the number of depth tokens varies between $[800, 1200]$ to reflect the input sparsity. We train the model using the AdamW optimizer with learning rates of $1 \times 10^{-5}, 1 \times 10^{-4}$ for the encoder and the decoder, respectively, halving the learning rate every 25k iterations. During the first 50k iterations, we adopt a low-resolution warm-up stage to stabilize training. For all evaluations, the final model is trained for 100k iterations on 8 NVIDIA A100 GPUs.

4.2. Evaluation Setting

We evaluate our approach under a strict zero-shot generalization setting using both real-sensor and simulated benchmarks. For real-sensor evaluation, we use KITTI-DC [32], captured with a Velodyne LiDAR, and ZJUL5 [46], obtained from the lightweight VL53L5CX sensor. These datasets present extremely sparse inputs (down to 8×8), span both indoor and outdoor scenes, and represent two distinct sensor families.

For simulated benchmarks, we randomly sample 500 sparse depth points from the ground-truth depth maps. We use DDAD [8] for outdoor driving scenes, DIODE [33] and ETH3D [29] for mixed indoor–outdoor scenarios, and iBims-1 [16] for indoor evaluations. As these datasets provide real sensor measurements, we follow the common depth completion protocol that generates sparser inputs via downsampling. Notably, DDAD is captured using Luminar-H2, adding another real dToF sensor under evaluation.

We compare against state-of-the-art depth completion and enhancement approaches, including OMNI-DC [50], Marigold-DC [34], PromptDA [19], and Prior Depth Anything [41]. As PromptDA does not natively support sparse depth, we complete missing regions with a flood-fill operation after nearest-neighbor downsampling to match a ViT-compatible resolution. Performance is reported using the relative error $\mathrm{rel} = \|\tilde{\mathbf{d}} - \mathbf{d}\|_1 / \mathbf{d}$ and the threshold accuracy $\delta = \max\left(\tilde{\mathbf{d}}/\mathbf{d},\, \mathbf{d}/\tilde{\mathbf{d}}\right) < 1.025$.

4.3. Quantitative Evaluation

Depth completion. Table 1 shows that our method achieves competitive or state-of-the-art results across all bench-

Table 1. The comparison of zero-shot generalization across real and simulated sparse-depth inputs. Each dataset block reports the relative depth error (Rel) and threshold accuracy (δ). The best, second-best, and third-best results are highlighted in dark, medium, and light green.

Input	Real sensor depth				Simulated points											
Method	KITTI-DC		ZJUL5		DDAD		DIODE		ETH3D		iBims-1		Average		Inference Cost	
	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	time ↓	memory↓
OMNI-DC	1.48	90.1	12.92	29.8	5.15	72.7	0.83	97.1	1.05	91.6	1.16	92.5	3.77	79.0	638 ms	3.49 GB
PromptDA(ViT-S)	2.90	72.6	13.25	29.4	7.20	49.4	2.11	88.7	2.67	76.7	2.42	81.2	5.09	66.3	30 ms	0.40 GB
PromptDA(ViT-L)	3.38	65.1	12.8	30.1	7.32	44.7	2.27	88.1	2.71	78.2	2.37	83.3	5.14	64.9	102 ms	1.56 GB
PriorDA(ViT-S)	2.85	75.2	11.53	32.1	6.36	60.5	1.27	92.9	1.30	89.6	1.52	88.7	4.14	73.2	118 ms	0.77 GB
PriorDA(ViT-B)	2.76	76.3	11.13	32.2	6.50	60.9	1.15	93.9	1.04	92.9	1.35	90.6	3.99	74.5	197 ms	1.62 GB
Marigold-DC	5.50	43.8	12.77	29.4	12.01	29.8	2.58	82.3	1.80	83.4	2.00	85.8	6.11	59.1	6470 ms	5.71 GB
Ours	2.00	84.0	11.13	32.3	4.61	68.8	0.89	96.7	0.84	94.8	1.29	90.9	3.46	77.9	34 ms	0.44 GB

Table 2. The comparison of zero-shot generalization under varying noise patterns: spatial shifting (*spa.*), precision degradation (*deg.*) and missing holes (*mis.*). Each dataset block shows relative error (Rel) of metric depth in percentage. The best, second-best, and third-best results are highlighted in dark, medium, and light green, respectively.

Method	DDAD			DIODE			ETH3D			iBims-1			Average		
	spa.	*deg.*	*mis.*	*spa.*	*deg.*	*mis.*	*spa.*	*deg.*	*mis.*	*spa.*	*deg.*	*mis.*	*spa.*	*deg.*	*mis.*
OMNI-DC	6.84	5.28	6.66	1.16	1.24	1.09	1.60	1.53	1.62	1.54	1.67	1.35	2.78	2.43	2.68
PromptDA	8.50	7.27	10.1	2.21	2.23	2.56	2.87	2.85	3.41	2.50	2.62	2.88	4.02	3.74	4.73
PriorDA	7.26	6.41	7.10	1.42	1.44	1.40	1.66	1.56	1.47	1.71	1.75	1.63	3.01	2.79	2.90
Marigold-DC	13.70	12.05	13.34	3.16	2.76	2.66	2.25	2.04	1.90	2.49	2.57	2.08	5.40	4.97	4.99
Ours	5.11	4.64	5.37	0.97	0.97	1.04	1.01	1.01	1.02	1.39	1.48	1.48	2.12	2.02	2.23

Table 3. Ablation study on the network architecture. We evaluate against decoder-level depth prompting following PromptDA, and joint attention without a directional masking. We also scale the network backbone to evaluate performance across model capacities. We report the average relative depth error (Rel) and threshold accuracy (δ). Refer to supplements for the full table.

Methods	Rel	δ
Depth prompting	4.07	73.6
Joint attention	3.87	75.6
Masked joint attention (Ours, ViT-S)	3.46	77.9
Ours, ViT-B	3.28	79.0
Ours, ViT-L	3.07	80.1

marks while maintaining high computational efficiency. Notably, our model achieves the best performance on ETH3D and ZJUL5. ETH3D includes high-resolution DSLR imagery (2048×1365), while ZJUL5 provides extremely low-resolution dToF measurements. These results highlight our model's ability to perform accurate metric depth reconstruction even when the input depth is vastly sparser than the image resolution.

We also compare inference speed and memory usage on an NVIDIA A100 GPU. PromptDA and our model are evaluated in FP16, while other methods follow their official settings. All timings are measured with batch size 1 and averaged across benchmarks at native input resolutions. Please refer to the supplementary material for the benchmark resolution specifications. Compared to OMNI-DC, our model runs 20× faster and requires 10× less memory, while achieving comparable or higher accuracy—demonstrating the strong efficiency–accuracy trade-off offered by our approach.

Robustness to noise. We further assess robustness to three common depth degradation types: spatial shifting, precision degradation, and missing-depth holes. Experiments are conducted on DDAD, DIODE, ETH3D, and iBims-1: (1) Spatial shifting: 500 points are randomly sampled and perturbed by spatial jitter up to 0.5% of the image diagonal. (2) Precision degradation: we introduce edge distortions via random erosion/dilation of the ground-truth depth, with erosion thickness set to 1% of the image diagonal, followed by sampling 500 points. (3) Missing holes: we remove depth regions using square masks or irregular Perlin-noise masks covering $20 \pm 5\%$ of the image.

We evaluate using relative error, consistent with the main experiments. As shown in Table 2, our method achieves the best results across all noise types, indicating strong resilience to depth acquisition artifacts.

Robustness to depth sparsity. To analyze performance under varying sparsity, we sample 100–2.5k points from ground-truth depth of DDAD, DIODE, ETH3D, and iBims-1. Relative errors are averaged across datasets and visualized in Figure 6. Our method achieves the highest accuracy in the extreme sparse regime (100 points) and exhibits the smallest performance drop as the number of points increases, demonstrating strong robustness to input density variations.

4.4. Qualitative Comparison

Figure 7 presents qualitative comparisons. For each method, we visualize the predicted metric depth, and the reconstructed 3D point cloud. Our method consistently recovers more accurate global geometry, particularly in regions with extremely sparse or missing depth. In the Library scene (row 1), competing methods significantly misestimate the ceiling distance, while our model reconstructs the correct global structure. In the Pipes scene (row 2), competing methods estimate distorted pipes, whereas our approach preserves a correct geometry. Our approach preserves correct spa-

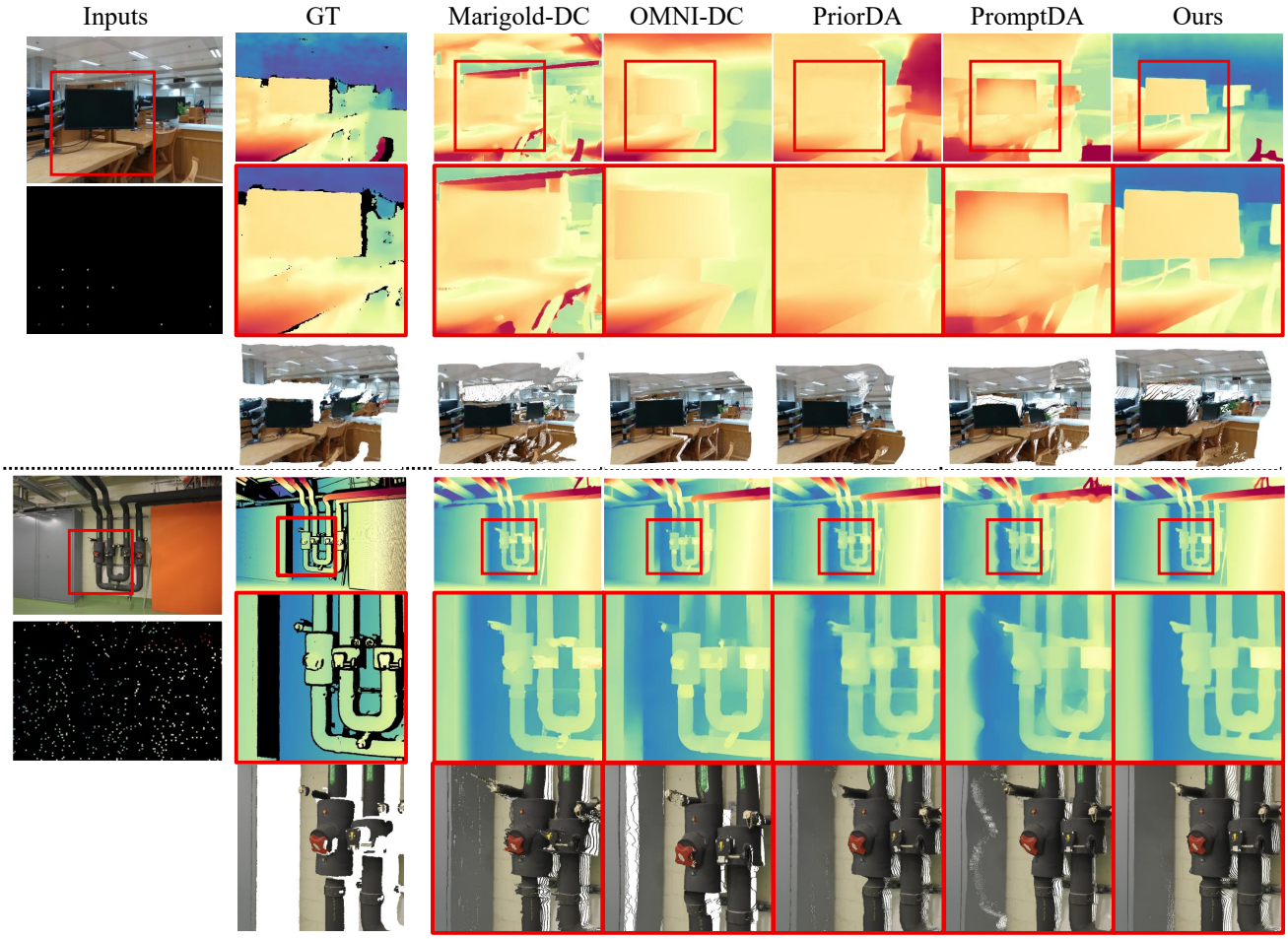

Figure 7. Qualitative comparison of metric depth maps, and reconstructed point clouds. The two examples are taken from the Library scene of ZJUL5, and the Pipes scenes of ETH3D.

tial relationships of local geometry and even for far-range structures with few depth observations.

4.5. Ablation study

Depth fusion architecture. We validate the importance of masked joint attention through two baselines: (1) a depth-prompting variant that follows PromptDA by injecting depth features only into the decoder's convolutional neck, and (2) joint attention without directional masking. All variants are trained under identical settings. As shown in Table 3, encoder-level fusion already outperforms decoder-level prompting, and introducing the directional mask further improves accuracy. These results confirm that masked joint attention is a crucial architectural component.

ViT backbone. We also report the results of our model using a ViT-Base and a ViT-Large backbone in Table 3. The performance further improves with the larger ViT model, demonstrating the scalability of our model. Note, our ViT-Small model already outperforms PromptDA and PriorDA that rely on larger ViT-B or ViT-L backbones. Our method is model size-agnostic by design, and we focus on ViT-Small to emphasize efficiency and practical deployment.

5. Conclusion

We presented a lightweight and generalizable framework for dense metric depth completion from sparse dToF measurements. Our depth-guided dual-branch encoder enables effective RGB and sparse depth fusion, while our dToF simulation pipeline greatly improves robustness across devices, sparsity levels, and noise conditions. Experiments on real and simulated benchmarks show that our method achieves strong accuracy and efficiency, establishing a practical solution for dense depth estimation from sparse dToF inputs.

Acknowledgements

Min H. Kim acknowledges the Samsung Research Funding & Incubation Center (SRFC-IT2402-02), the Korea NRF grant (RS-2024-00357548), the MSIT/IITP of Korea (RS-2022-00155620, RS-2024-00398830, RS-2024-00436680, and 2017-0-00072), the MSIT Advanced GPU Utilization Support Program, and Microsoft Research Asia.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 5

[2] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations*. 2

[3] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE transactions on pattern analysis and machine intelligence*, 42 (10):2361–2379, 2019. 2

[4] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 5

[5] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 5

[6] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 2

[7] Jose L Gómez, Manuel Silva, Antonio Seoane, Agnès Borrás, Mario Noriega, Germán Ros, Jose A Iglesias-Guitian, and Antonio M López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *Neurocomputing*, 637:130038, 2025. 5

[8] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[9] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9233–9243, 2023. 2

[10] Yuliang Guo, Guang Chen, Peitao Zhao, Weide Zhang, Jinghao Miao, Jingao Wang, and Tae Eun Choe. Gen-lanenet: A generalized and scalable approach for 3d lane detection. In *European Conference on Computer Vision*, pages 666–681. Springer, 2020. 5

[11] Jing He, LI Haodong, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *The Thirteenth International Conference on Learning Representations*. 2

[12] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13656–13662. IEEE, 2021. 2

[13] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[14] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2821–2830, 2018. 5

[15] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9492–9502, 2024. 2

[16] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 6

[17] Hoang-An Le, Partha Das, Thomas Mensink, Sezer Karaoglu, and Theo Gevers. EDEN: Multimodal Synthetic Dataset of Enclosed garDEN Scenes. In *Proceedings of the IEEE/CVF Winter Conference of Applications on Computer Vision (WACV)*, 2021. 5

[18] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3215, 2023. 5

[19] Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17070–17080, 2025. 2, 6

[20] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[21] Zhiheng Liu, Ka Leong Cheng, Qiuyu Wang, Shuzhe Wang, Hao Ouyang, Bin Tan, Kai Zhu, Yujun Shen, Qifeng Chen, and Ping Luo. Depthlab: From partial to complete. *arXiv preprint arXiv:2412.18153*, 2024. 2, 3, 5

[22] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics (ToG)*, 38(6):1–15, 2019. 5

[23] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Di-

nov2: Learning robust visual features without supervision, 2023. 3, 4

[24] Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9869, 2024. 2

[25] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 2

[26] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 4, 5

[27] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021.

[28] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 5

[29] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[30] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9763–9772, 2024. 2

[31] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[32] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 6

[33] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR*, abs/1908.00463, 2019. 6

[34] Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5359–5370, 2025. 2, 3, 5, 6

[35] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020. 5

[36] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019. 5

[37] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 2, 4, 5

[38] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 2, 6

[39] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020. 5

[40] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9422–9432, 2023. 2

[41] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. *arXiv preprint arXiv:2505.10565*, 2025. 2, 3, 6

[42] Magnus Wrenninge and Jonas Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing. *arXiv preprint arXiv:1810.08705*, 2018. 5

[43] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10371–10381, 2024. 2

[44] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37: 21875–21911, 2024. 2

[45] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5

[46] Li Yijin, Liu Xinyang, Dong Wenqi, Zhou han, Bao Hujun, Zhang Guofeng, Zhang Yinda, and Cui Zhaopeng. Deltar: Depth estimation from a light-weight tof sensor and rgb image. 2022. 6

[47] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9043–9053, 2023. 2

[48] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pages 519–535. Springer, 2020. 5

[49] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *European Conference on Computer Vision*, pages 78–95. Springer, 2024. 3

[50] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9297, 2025. 2, 3, 6