# GEOMETRY-AWARE METRIC SCALE PROPAGATION FOR ROBUST DEPTH COMPLETION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Depth completion aims to recover a dense depth map from an RGB image and sparse depth measurements. Recent depth completion methods leverage depth foundation models to substantially improve real-world generalization. However, how to fuse the crucial metric scale information from sparse depth measurements with relative depth remains a challenge. In this paper, we identify that sparse depth patterns in the 2D plane can lead to scale ambiguity for depth completion in the 3D scene. That is, similar sparse depth patterns may correspond to different 3D geometries, resulting in ambiguous scale correspondence between sparse measurements and relative depth. It consequently yields distorted 3D geometry in depth completion. To mitigate, we propose *GeoSP*, a *Geo*metry-aware *S*cale *P*ropagation framework for robust depth completion, which propagates metric scale from sparse depth to dense relative depth guided by 3D geometry. More specifically, GeoSP identifies 3D geometric relations between sparse depth measurements and dense relative depth, and then formulates them as propagation cues to compute metric scale in depth completion. It contains three complementary modules. *Geometric Neighborhood Selection* uses the relative 3D geometry priors to select valid scale constraints and identify geometric relations. *Geometric Affinity Prediction* then introduces a lightweight head with ground-truth 3D geometry supervision to adaptively estimate the contribution of each constraint from these relations. Finally, *Affinity-Weighted Solver* integrates the selected valid scale constraints according to their predicted affinities to compute metric scale. Extensive experiments are conducted on ten unseen benchmarks with diverse sparse depth patterns. Our GeoSP achieves state-of-the-art performance with lower complexity, particularly in 3D geometric accuracy, when compared to both traditional models and recent depth-foundation-model-based methods.

## 1 INTRODUCTION

Depth completion aims to recover dense depth from an RGB image and sparse depth measurements, providing essential geometric information for 3D perception. Traditional methods Ma & Karaman (2018); Cheng et al. (2018); Park et al. (2020); Zhang et al. (2023) achieve impressive results on benchmarks such as NYU Depth V2 Silberman et al. (2012) and KITTI Geiger et al. (2012), however, they typically exhibit a heavy reliance on in-domain RGB-D supervision. Consequently, their performance degrades significantly when generalizing to real-world scenes.

Recently, relative depth foundation models such as MiDaS (Ranftl et al., 2022) and Depth Anything (Yang et al., 2024a;b) have demonstrated strong generalization to unseen scenes. These models can predict dense relative depth from single RGB images. Building on this capability, recent depth completion methods (Wang et al., 2026; Chen et al., 2025; Zhou et al., 2026) leverage these models to improve real-world generalization. A representative example is PSD (Chen et al., 2025), which uses the predicted relative depth as a scene prior to recover dense depth in missing regions from sparse measurements.

Although relative depth already provides dense scene geometry, it lacks the crucial metric scale information that is required to convert a dimensionless relative depth into physical distances. In depth completion, sparse depth measurements provide such metric scale information, but only at a



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

limited set of locations. Therefore, how to fuse the crucial metric scale information from sparse depth measurements with dense relative depth remains a challenge.

In this paper, we identify that sparse depth patterns in the 2D plane can lead to scale ambiguity for depth completion in the 3D scene. That is, similar sparse depth patterns may correspond to different 3D geometries, resulting in different scale correspondences between sparse measurements and relative depth. Such ambiguity can yield incorrect metric depth, which consequently manifests as distorted 3D geometry.

To mitigate, we propose *GeoSP*, Geometry-Aware Scale Propagation for Robust Depth Completion. GeoSP leverages a frozen depth foundation model to provide dense relative depth and propagates metric scale from sparse depth to dense relative depth guided by 3D geometry. More specifically, GeoSP first identifies 3D geometric relations between sparse depth measurements and dense relative depth to mitigate the scale ambiguity from 2D sparse patterns. These relations are then formulated as scale propagation cues to explicitly compute metric scale.

Specifically, GeoSP contains three complementary modules. *Geometric Neighborhood Selection* first uses the relative 3D geometry prior to select sparse measurements that provide valid scale constraints, and then identifies their geometric relations to the dense relative geometry. *Geometric Affinity Prediction* then introduces a lightweight head supervised by ground-truth 3D geometry to adaptively estimate the contribution of each constraint from these geometric relations, helping reduce the influence of potentially unreliable geometry. Finally, *Affinity-Weighted Solver* explicitly integrates the selected scale constraints according to their predicted affinities and solves for the local metric scale, thereby transferring reliable metric information from sparse measurements to the dense relative depth and producing recovered dense depth.

Our GeoSP is trained only on a mixture of NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012), and extensively evaluated on ten unseen benchmarks with diverse simulated and raw sparse depth patterns. Compared with both traditional and recent depth-foundation-model-based methods, GeoSP achieves state-of-the-art overall performance, with particularly higher 3D geometric accuracy while remaining competitive in 2D depth accuracy. Notably, these results are achieved while requiring only a lightweight trainable head.

Our main contributions are highlighted as follows.

- We identify that the patterns of sparse depth measurements in the 2D plane can lead to the ambiguity of the crucial scale information in depth completion in the 3D scene.
- We propose a geometry-aware scale propagation framework for robust depth completion. It propagates metric scale from sparse depth to relative depth guided by 3D geometry.
- Our GeoSP achieves state-of-the-art performance on average especially in terms of 3D geometries, when compared to both traditional and recent foundation-model-based methods.

## 2 Related Work

### 2.1 Traditional Depth Completion

Depth completion has achieved remarkable progress in recent years. Traditional methods mainly focus on improving prediction accuracy under in-domain settings. Early learning-based methods directly combine RGB images and sparse depth to predict dense depth (Ma & Karaman, 2018). Later methods progressively improve how RGB information and sparse depth are used together, leading to more accurate depth completion (Qiu et al., 2019; Tang et al., 2021; Liu et al., 2021; Zhao et al., 2021; Yan et al., 2022; Zhang et al., 2023). More recent propagation-based methods explicitly spread sparse depth constraints to missing regions through learned affinities (Cheng et al., 2018; Park et al., 2020; Lin et al., 2024; Tang et al., 2024).

These methods achieve impressive results on benchmarks such as NYU Depth V2 and KITTI. Nonetheless, they typically exhibit a heavy reliance on in-domain RGB-D supervision. Consequently, the learned completion and propagation rules often generalize poorly to unseen scenes or different sparse depth patterns.

### 2.2 Relative Depth Foundation Models

Foundation models for monocular depth estimation have been trained on large-scale and diverse datasets, showing strong generalization to unseen scenes. MiDaS (Ranftl et al., 2022) introduces a scale-invariant training objective to learn relative depth across diverse scenes. Depth Anything (Yang et al., 2024a) further scales up training through a data engine for large-scale automatic annotation, while Depth Anything V2 (Yang et al., 2024b) improves depth estimation by using high-quality synthetic data and refined training strategies. Marigold (Ke et al., 2024) demonstrates that generative image priors can also be repurposed for robust affine-invariant depth estimation. More recently, MoGe (Wang et al., 2024b) extends monocular geometry estimation from depth prediction to dense 3D point prediction, producing an affine-invariant point map.

Recent depth completion methods leverage relative depth as prior to improve real-world generalization. However, how to effectively fuse crucial metric scale information from sparse depth measurements with relative depth remains a challenge.

### 2.3 Foundation-Model-Based Depth Completion

To recover the metric scale of relative depth predicted by a depth foundation model, most existing methods establish correspondences between sparse metric measurements and relative depth in the 2D image plane. PriorDA (Wang et al., 2026) expands sparse measurements into a coarse metric prior, while PromptDA (Lin et al., 2025), Any2Full (Zhou et al., 2026), and DepthPrompting (Park et al., 2024) incorporate sparse metric information through features, prompts, or image-domain propagation. TestPromptDC (Jeong et al., 2025) and Marigold-DC (Viola et al., 2025) instead exploit sparse measurements directly at test time. PSD (Chen et al., 2025) simply uses 3D Euclidean distance only for selecting propagation neighbors, but its geometric information is mainly used for neighborhood construction, followed by a learned depth correction stage.

In contrast, GeoSP explicitly identifies 3D geometric relations and formulates them as constraints for computing local metric scale, allowing both valid scale constraints and their contributions to be determined during scale propagation.

## 3 Method

### 3.1 Overview

The proposed framework is illustrated in Fig. 1. Given an RGB image $I$ and corresponding sparse depth measurements $D_s$, our goal is to recover a dense depth map $D^\star \in \mathbb{R}^{H\times W}$. We first leverage a frozen depth foundation model (Yang et al., 2024b) to predict relative depth map $D_m \in \mathbb{R}^{H\times W}$. We therefore perform geometry-aware scale propagation to recover its local metric scale. To identify the 3D geometric relations between sparse depth measurements and relative depth, we first lift them into a shared 3D space. Since the relative depth $D_m$ lacks metric scale, we first perform *Global Alignment* with the sparse depth measurements $D_s$ by estimating a global scale $\gamma$ and shift $\rho$, obtaining

$$\widetilde{D}_m = \gamma D_m + \rho. \tag{1}$$

We then back-project the aligned relative depth $\widetilde{D}_m$ and sparse depth $D_s$ using the camera intrinsic matrix $K$ into a shared 3D, where their geometric relations can be explicitly established for subsequent geometry-aware scale propagation.

Building on this, GeoSP decomposes geometry-aware scale propagation into three sequential modules. *Geometric Neighborhood Selection* first selects a sparse neighborhood $\mathcal{N}$ as valid scale constraints in the shared 3D space and establishes the geometric relations $\phi$ between the selected constraints and the dense relative geometry. *Geometry Affinity Prediction* then takes the selected sparse constraints $\mathcal{N}$ and their geometric relations $\phi$ as input to estimate their contributions as affinities $A$. Finally, the *Affinity-Weighted Solver* integrates the selected constraints $\mathcal{N}$ according to $A$ to explicitly propagate metric scale to relative depth, yielding the dense depth ($D^\star$).

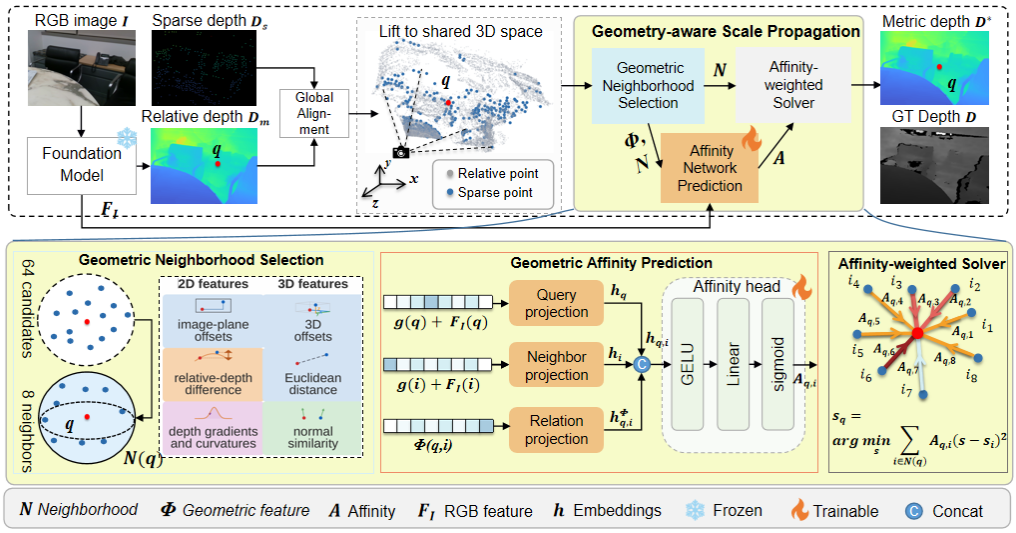



Figure 1: Overview of GeoSP. Given an RGB image $I$ and sparse depth $D_s$, a frozen depth foundation model first predicts relative depth $D_m$. The relative and sparse depth are then lifted into a shared 3D space, where their geometric relations are explicitly established for scale propagation. GeoSP then propagates the metric scale carried by sparse measurements to the dense relative depth through three sequential modules: *Geometric Neighborhood Selection* identifies valid sparse scale constraints for relative depth, *Geometric Affinity Prediction* estimates their contributions, and the *Affinity-Weighted Solver* integrates the selected constraints according to their predicted affinities to explicitly recover the local metric scale.

### 3.2 Geometric Neighborhood Selection.

Following the overview, we detail how our method selects sparse measurements as scale constraints and characterizes their geometric relations. For each dense query $q$ of relative depth, we first collect the 64 spatially nearest sparse measurements in the image plane as a local candidate pool. We then measure the 3D Euclidean distance between the query and each candidate in the shared 3D space and select the $k$=8 nearest measurements to form the geometric neighborhood $\mathcal{N}(q)$.

For each selected neighbor $i \in \mathcal{N}(q)$, we further identify pairwise geometric relations $\phi(q, i)$ to characterize its relation to the query. The relation includes 2D positional cues, such as image-plane positions and offsets; depth-related cues, including relative-depth differences and local depth gradients and curvatures; and 3D geometric cues, including 3D offsets, Euclidean distance, and normal similarity. Together, the selected sparse neighborhood $\mathcal{N}(q)$ and the corresponding pairwise relations $\phi(q, i)$ provide the sparse scale constraints and their geometric context for the next module, which further estimates the contribution of each constraint to local metric scale recovery.

### 3.3 Geometric Affinity Prediction

Given the geometric neighborhood $\mathcal{N}(q)$ and the pairwise relations $\phi(q, i)$ established by the previous module, our method further determines how strongly each selected constraint should contribute to local metric scale estimation. Specifically, we predict an affinity $A_{q,i}$ for each $i \in \mathcal{N}(q)$, where a larger affinity indicates a stronger contribution of the corresponding sparse measurement to the metric scale recovery at $q$.

To estimate this contribution, we characterize not only the geometric relations between the query and each selected constraint, but also the properties of the two points themselves. Each point is represented by point-wise geometric cues $g(\cdot)$, including its 3D position, surface normal, and depth. Appearance features $F_I$ are extracted from the frozen depth foundation model and sampled at the corresponding locations. We then represent the query and neighboring measurement as $[F_I(q), g(q)]$ and $[F_I(i), g(i)]$, respectively, while their pairwise geometric relations are represented by $\phi(q, i)$ obtained from the neighborhood selection module.

To jointly model the query, the sparse constraint, and their geometric relations for affinity estimation, three projection heads map these representations into a shared 128-dimensional embedding space:

$$\mathbf{h}_q = P_q\big([F_I(q), g(q)]\big), \quad \mathbf{h}_i = P_i\big([F_I(i), g(i)]\big), \quad \mathbf{h}^{\phi}_{q,i} = P_e\big(\phi(q,i)\big), \tag{2}$$

where $P_q$, $P_i$, and $P_e$ denote the query, neighbor, and relation projections, respectively. The query embedding $\mathbf{h}_q$ is computed once and shared across all $i \in \mathcal{N}(q)$. The three embeddings are concatenated as

$$\mathbf{h}_{q,i} = \mathrm{Concat}\left(\mathbf{h}_q, \mathbf{h}_i, \mathbf{h}^{\phi}_{q,i}\right), \tag{3}$$

and a lightweight prediction head maps the fused representation to an affinity,

$$a_{q,i} = \mathrm{Head}(\mathbf{h}_{q,i}), \qquad A_{q,i} = \sigma(a_{q,i}), \tag{4}$$

where the prediction head consists of a GELU activation followed by a linear layer, and $\sigma$ denotes the sigmoid function.

The predicted affinities are then used as weights in the subsequent affinity-weighted least-squares solver to explicitly recover the local metric scale. Importantly, the network does not directly predict metric depth or scale, but only the propagation weights of the sparse constraints. The affinities are therefore learned indirectly through supervision on the recovered metric depth, encouraging the predicted weights to favor constraints that support geometrically consistent scale recovery. The metric scale itself is explicitly computed by the closed-form solver described next.

### 3.4 Affinity-weighted solver

Our method recovers the local metric scale from the selected sparse constraints using their predicted affinities. Given the selected scale constraints $\mathcal{N}(q)$ and their predicted affinities $A_{q,i}$, each sparse measurement $i \in \mathcal{N}(q)$ provides a metric scale constraint by comparing its metric depth $D_s(i)$ with the aligned relative depth $\widetilde{D}_m(i)$ at the same location:

$$s_i = \frac{D_s(i)}{\widetilde{D}_m(i)}. \tag{5}$$

Here, $s_i$ specifies the metric scale of sparse depth measurement $i$.

We then recover the local metric scale $s_q$ at query $q$ by integrating the selected scale constraints according to their predicted affinities:

$$s_q = \arg\min_{s} \sum_{i \in \mathcal{N}(q)} A_{q,i} \left(s - s_i\right)^2. \tag{6}$$

This affinity-weighted least-squares objective admits a closed-form solution. The recovered scale is then applied to the relative depth as $D^{\star}(q) = s_q \widetilde{D}_m(q)$. Applying this process to all dense queries propagates scale over the dense relative geometry and yields the dense depth $D^{\star}$.

### 3.5 Training Supervision

We freeze the depth foundation model and train only the affinity network. Supervision is applied to the recovered metric depth $D^{\star}$, so the affinity head is optimized indirectly.

**Depth supervision.** $\mathcal{L}_{\mathrm{pix}}$ applies an $\ell_1$ loss in log-depth space over all valid pixels (Eigen et al., 2014; Ranftl et al., 2022). $\mathcal{L}_{\mathrm{low}}$ applies the same loss after Gaussian smoothing to constrain the smooth metric depth variations across the image, following multi-scale log-depth supervision (Eigen et al., 2014). $\mathcal{L}_{\mathrm{sp}}$ applies the same log-space $\ell_1$ loss at sparse measurement locations (Ma & Karaman, 2018), but on the unpinned reconstruction before $D_s$ is directly enforced. The depth terms are combined as

$$\mathcal{L}_{\mathrm{depth}} = \mathcal{L}_{\mathrm{pix}} + \mathcal{L}_{\mathrm{low}} + \lambda_{\mathrm{s}} \mathcal{L}_{\mathrm{sp}}. \tag{7}$$

**3D Geometric supervision.** $\mathcal{L}_{\mathrm{pt}}$ measures the scene-normalized $\ell_1$ distance between corresponding predicted and ground-truth 3D points (Wang et al., 2024b). $\mathcal{L}_{\mathrm{pl}}$ further constrains local surface geometry through point-to-plane residuals along ground-truth normals. These terms constrain the back-projected point map $X$ of $D^\star$:

$$\mathcal{L}_{\mathrm{3D}} = \mathcal{L}_{\mathrm{pt}} + \lambda_{\mathrm{pl}}\mathcal{L}_{\mathrm{pl}}. \tag{8}$$

Ground-truth 3D geometry thus supervises the affinity predictor, helping reduce the influence of potentially unreliable geometric relations from the relative depth.

The full training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{3D}}\mathcal{L}_{\mathrm{3D}}. \tag{9}$$

# 4 Experiments

## 4.1 Experimental Setup

**Training datasets.** Our framework is trained on a mixture of NYUv2 (Silberman et al., 2012) and KITTI (Geiger et al., 2012), following the training splits of (Bhat et al., 2023). The training set contains 47,388 RGB-D frames, including 24,231 frames from NYUv2 and 23,157 frames from the KITTI Eigen (Eigen et al., 2014) split. Importantly, all sparse depth used for training is simulated from ground-truth depth, and no raw sensor sparse depth is observed during training. We simulate LiDAR patterns with 4/8/16/32 lines and random sampling with 100–2000 points, covering substantial variations in both observation density and spatial pattern. This training strategy exposes GeoSP to diverse sparse settings while avoiding specialization to any particular sensing pattern.

**Evaluation settings.** We evaluate our method under two representative settings:

(i) *Simulated sparse depth*. We evaluate GeoSP on three indoor and three outdoor benchmarks, covering diverse scenes with substantially different sparse-depth patterns. On indoor benchmarks, we use random sampling at 0.1% and 1% of valid depth points, together with Harris-corner and hole patterns. On outdoor benchmarks, random sampling is replaced by simulated 8-line and 16-line LiDAR observations, while Harris-corner and hole patterns are retained. Harris-corner sampling follows Chen et al. (2025), and the hole pattern follows Wang et al. (2024a).

(ii) *Raw sparse depth*. GeoSP is trained exclusively with simulated sparse depth and never observes raw sensor measurements during training. We therefore directly evaluate its zero-shot generalization to raw sparse depth captured by different real-world sensors. We follow the raw-depth protocol of (Chen et al., 2025) on VOID (Wong et al., 2020), SUN RGB-D (Song et al., 2015), DIML indoor (Kim et al., 2018), and TOFDC (Yan et al., 2024), directly using the provided sensor-native sparse measurements without additional resampling or target-domain adaptation. These benchmarks cover diverse sensing mechanisms, including VIO keypoints, incomplete RGB-D measurements, structured-light depth, and low-resolution ToF depth.

**Evaluation metrics.** To evaluate both depth accuracy and the 3D geometric accuracy of the recovered dense depth, we report metrics in both the 2D image domain and 3D space. For 2D depth, we use mean absolute error (MAE, mm), which directly measures the absolute metric depth error between the prediction and ground truth over valid pixels and thus provides an intuitive measure of metric depth accuracy. For 3D geometric quality, we further report Chamfer overall distance (Ov., mm) and the 50,mm F-score (F), computed between point clouds back-projected from the predicted and ground-truth depth maps. All camera intrinsics used for back-projection are provided by the corresponding datasets. Finally, we report the average rank (Rank) across all datasets and metrics to summarize overall performance, together with the 3D average rank (3D Rank) computed over Ov. and F to summarize geometric performance.

**Baselines.** We compare GeoSP with ten representative approaches. methods include NLSPN (Park et al., 2020), CompletionFormer (Zhang et al., 2023), and G2-MD (Wang et al., 2024a). Foundation-Model-Based methods include PSD (Chen et al., 2025), PriorDA (Wang et al., 2026), Any2Full (Zhou et al., 2026), DepthPrompting (Park et al., 2024), PromptDA (Lin et al., 2025), TestPromptDC (Jeong et al., 2025), and Marigold-DC (Viola et al., 2025).

Table 1: Comparison on simulated sparse depth. "FM" denotes the use of a depth foundation model. "2D" evaluates metric depth accuracy, while "3D" evaluates 3D geometric accuracy. "Avg. Rank" denotes the average ranking across the corresponding evaluation metrics. Best and second-best results in each column are highlighted in bold and underline, respectively.

| Method | FM | DIODE Indoor | | | Stanford | | | Middlebury | | | Average Rank | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2D | 3D | | 2D | 3D | | 2D | 3D | | All↓ | 3D↓ |
| | | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | | |
| CompletionFormer (Zhang et al., 2023) | – | 691 | 209 | 47.4 | 176 | 199 | 16.9 | 1165 | 233 | 35.7 | 8.66 | 8.50 |
| NLSPN (Park et al., 2020) | – | 518 | 161 | 58.7 | 124 | 178 | 21.8 | 783 | 162 | 47.9 | 5.66 | 5.00 |
| G2-MD (Wang et al., 2024a) | – | 574.9 | 282.9 | 43.0 | **83.89** | 166.6 | 20.0 | 1465 | 669.2 | 31.0 | 7.28 | 7.75 |
| DepthPrompting (Park et al., 2024) | ✓ | 398 | 173 | 58.0 | 138 | 185 | 20.0 | 700 | 162 | 46.0 | 6.22 | 6.17 |
| Marigold-DC (Viola et al., 2025) | ✓ | 3581 | 1321 | 37.8 | 223 | 287 | 13.1 | 280 | 109 | 44.6 | 8.56 | 8.84 |
| PromptDA (Lin et al., 2025) | ✓ | 932 | 257 | 36.0 | 369 | 309 | 7.00 | 1146 | 286 | 25.0 | 10.11 | 10.34 |
| PriorDA (Wang et al., 2026) | ✓ | 395 | 174 | 60.0 | 98.2 | 173 | 19.0 | **210** | 98.7 | 51.0 | 3.66 | 4.16 |
| Any2Full (Zhou et al., 2026) | ✓ | 4243 | 2173 | 52.0 | 104 | 178 | 19.0 | 324 | 139 | 48.0 | 7.23 | 7.17 |
| TestPromptDC (Jeong et al., 2025) | ✓ | 406 | 203 | 62.0 | 99.9 | 171 | 17.0 | 258 | 98.7 | 50.0 | 4.55 | 4.66 |
| PSD (Chen et al., 2025) | ✓ | 286 | 132 | 62.0 | 89.3 | 163 | **22.0** | 338 | 109 | 51.0 | 2.67 | 2.17 |
| **GeoSP (Ours)** | ✓ | **192** | **98.8** | **68.0** | 89.1 | **161** | 22.0 | 240 | **91.4** | **55.0** | **1.33** | **1.17** |

| Method | FM | DIML Outdoor | | | DIODE Outdoor | | | DrivingStereo | | | Average Rank | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2D | 3D | | 2D | 3D | | 2D | 3D | | All↓ | 3D↓ |
| | | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | | |
| CompletionFormer (Zhang et al., 2023) | – | 4469 | 933 | 9.00 | 4769 | 1339 | 6.40 | 4441 | 444 | 10.7 | 7.67 | 7.33 |
| NLSPN (Park et al., 2020) | – | 5766 | 874 | 10.4 | 4252 | 1008 | 6.68 | 3725 | 498 | 10.5 | 7.33 | 6.83 |
| G2-MD (Wang et al., 2024a) | – | 6643 | 1787 | 0.6 | 1862 | 572.6 | 5.0 | 1470 | 387.7 | 12.0 | 7.44 | 8.17 |
| DepthPrompting (Park et al., 2024) | ✓ | 4490 | 1340 | 5.00 | 3185 | 779 | 3.00 | 2258 | 348 | 11.0 | 7.44 | 7.67 |
| Marigold-DC (Viola et al., 2025) | ✓ | 6871 | 1146 | 1.00 | 3315 | 737 | 2.90 | 1524 | 288 | 13.3 | 7.50 | 7.25 |
| PromptDA (Lin et al., 2025) | ✓ | – | – | 1.00 | – | 5608 | 1.00 | – | – | 1.00 | 10.38 | 10.38 |
| PriorDA (Wang et al., 2026) | ✓ | 3489 | 1269 | 10.0 | 2200 | 538 | 8.00 | 1176 | 268 | 19.0 | 4.61 | 4.75 |
| Any2Full (Zhou et al., 2026) | ✓ | 3967 | 1406 | 15.0 | – | – | 6.00 | 883 | 223 | 20.0 | 4.86 | 5.20 |
| TestPromptDC (Jeong et al., 2025) | ✓ | 3704 | 812 | 4.00 | 2119 | 544 | 8.00 | **707** | **201** | 20.0 | 3.50 | 3.75 |
| PSD (Chen et al., 2025) | ✓ | 3472 | **583** | 16.0 | **1544** | **368** | 14.0 | 922 | 220 | 20.0 | 2.11 | 2.00 |
| **GeoSP (Ours)** | ✓ | **2714** | 761 | **19.0** | 1651 | 396 | **16.0** | 821 | 219 | **22.0** | **1.56** | **1.50** |

**Implementation details.** We use Depth Anything V2 Yang et al. (2024b) as the frozen monocular depth foundation model. Unlike PSD, we do not adopt a fixed-size cropping or resizing protocol, and instead train on NYUv2 at its native resolution of $640\times480$ and KITTI Eigen frames at their original resolution of approximately $1242\times375$. The model is optimized with AdamW for 3 epochs using an initial learning rate of $3\times10^{-4}$, which is decayed by a factor of $0.5$ after each epoch. Training samples are drawn with replacement using a 60/40 NYU–KITTI weighting, with 6,058 optimizer steps per epoch. The batch size is set to 1 with gradient accumulation over 2 steps. All experiments are conducted on a single GPU with 24 GB memory.

### 4.2 Main Results

**Results on simulated sparse Depth.** Table 1 reports the generalization comparison under simulated sparse depth on three indoor benchmarks, DIODE indoor (Vasiljevic et al., 2019), Stanford (Armeni et al., 2017), and Middlebury (Scharstein et al., 2014), and three outdoor benchmarks, DIML outdoor (Kim et al., 2018), DIODE outdoor, and DrivingStereo (Yang et al., 2019). Each reported result is averaged over four representative sparse patterns.

GeoSP achieves the best overall Avg. Rank on both indoor (**1.33**) and outdoor (**1.56**) benchmarks. The advantage is particularly evident in 3D geometry, where GeoSP obtains the best 3D Avg. Rank of **1.17** indoors and **1.50** outdoors. Across all six benchmarks, GeoSP further improves the F-score by **10.1%** on average while maintaining comparable Chamfer distance to the second best PSD. These results indicate that the gains in metric depth estimation are consistently accompanied by stronger 3D geometric accuracy, supporting the effectiveness of geometry-aware scale propagation across unseen scenes and diverse sparse patterns.

**Results on raw sparse Depth.** Table 2 reports the comparison under raw sparse depth settings using official sensor observations from DIML indoor Kim et al. (2018) with raw structured-light depth, SUN RGB-D Song et al. (2015) with incomplete RGB-D captures, TOFDC Yan et al. (2024)

Table 2: Comparison on raw sparse depth. GeoSP achieves the best overall Avg. Rank and 3D Avg. Rank.

| Method | FM | DIML Indoor | | | SUN RGB-D | | | TOFDC | | | VOID1500 | | | Average Rank | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2D | 3D | | 2D | 3D | | 2D | 3D | | 2D | 3D | | | |
| | | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | MAE↓ | Ov.↓ | F↑ | All↓ | 3D↓ |
| CompletionFormer (Zhang et al., 2023) | – | 468.9 | 141 | 26.43 | 478 | 270 | 9.58 | 41.5 | 32.0 | 89.5 | 118.8 | 57.8 | 71.8 | 7.17 | 6.94 |
| NLSPN (Park et al., 2020) | – | 510.1 | 149 | 26.05 | 520 | 280 | 9.59 | 41.5 | 32.0 | 89.5 | 80.9 | 43.8 | 83.0 | 7.29 | 7.00 |
| G2-MD (Wang et al., 2024a) | – | 432 | 140 | 26.9 | 5361 | 3155 | 0.542 | 40.3 | 31.7 | 89.6 | 35.3 | 26.6 | 89.6 | 6.38 | 6.19 |
| DepthPrompting (Park et al., 2024) | ✓ | 424 | 134 | 27.6 | 523 | 282 | 9.59 | 40.9 | 32.0 | 89.5 | 52.6 | 33.1 | 85.4 | 5.79 | 5.31 |
| Marigold-DC (Viola et al., 2025) | ✓ | 418 | 142 | 26.7 | 692 | 322 | 4.91 | 39.8 | 32.1 | 88.8 | 38.5 | 27.0 | 90.1 | 6.83 | 7.50 |
| PromptDA (Lin et al., 2025) | ✓ | 641 | 200 | 24.6 | 885 | 374 | 3.18 | 229 | 116 | 55.4 | 243 | 111 | 56.7 | 10.58 | 10.63 |
| PriorDA (Wang et al., 2026) | ✓ | **296** | **126** | **30.7** | 688 | 307 | 6.25 | 41.9 | 34.5 | 85.0 | 43.2 | 28.2 | 88.8 | 6.33 | 6.38 |
| Any2Full (Zhou et al., 2026) | ✓ | 375 | 129 | 27.4 | 2475 | 1235 | 7.38 | **38.9** | 31.7 | 89.7 | 36.8 | 25.9 | 90.3 | 4.21 | 4.06 |
| TestPromptDC (Jeong et al., 2025) | ✓ | 383 | 134 | 26.6 | 596 | 297 | 6.49 | 39.3 | 32.0 | 89.1 | **30.8** | **24.1** | 90.9 | 4.63 | 5.38 |
| PSD (Chen et al., 2025) | ✓ | 419 | 138 | 27.1 | 446 | 253 | 9.44 | 40.0 | 31.9 | 89.4 | 36.4 | 27.5 | 88.9 | 4.75 | 5.13 |
| **GeoSP (Ours)** | ✓ | 383 | 130 | 27.9 | **408** | **250** | **9.67** | 40.1 | **31.5** | **89.9** | 36.2 | 25.2 | **91.8** | **2.04** | **1.50** |

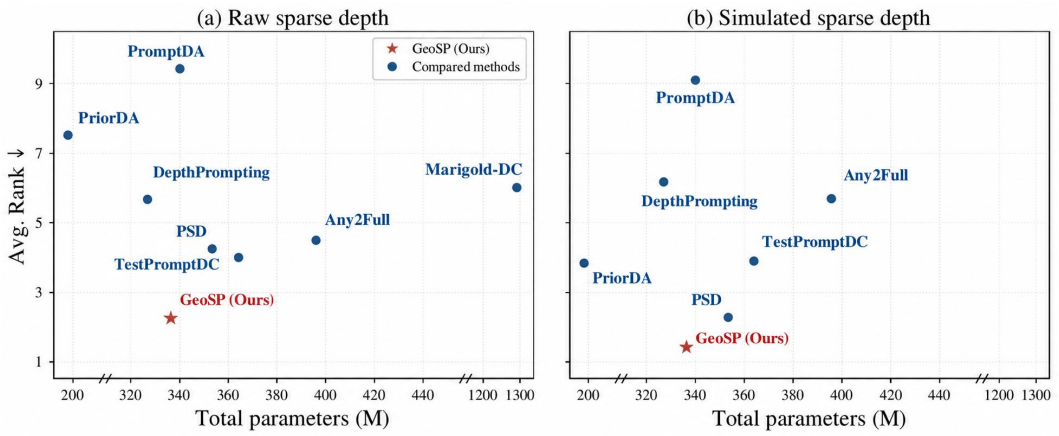



Figure 2: Accuracy–complexity trade-off. Our GeoSP achieves the best overall performance with relatively lower complexity.

with low-resolution ToF depth, and VOID Wong et al. (2020) with VIO-based sparse depth with 1500 points.

GeoSP achieves the best overall Avg. Rank of **2.04**, substantially outperforming the second-best Any2Full (4.21). More notably, GeoSP achieves the best 3D Avg. Rank of **1.50**, with a substantial margin over the second-best Any2Full (4.06). Compared with Any2Full, GeoSP reduces the 3D Chamfer distance by **20.6%** and improves the F-score by **8.7%** on average across the four benchmarks. Despite being trained only with simulated sparse depth patterns, GeoSP generalizes consistently to raw measurements from diverse real-world sensors, demonstrating strong sim-to-real generalization in both depth accuracy and 3D geometry.

**Accuracy and complexity trade-off.** Fig. 2 presents the Accuracy–complexity trade-off among foundation-model-based methods on raw sparse depth (left) and simulated sparse patterns (right). The Avg. Ranks are taken from Tables 1 and 2; for the simulated setting, the reported rank is the arithmetic mean of the indoor and outdoor Avg. Ranks. The horizontal axis reports the total number of model parameters used for inference. In both settings, GeoSP achieves the best Avg. Rank while maintaining a relatively small overall model size, with only DepthPrompting and PriorDA using fewer parameters among the compared methods. Moreover, GeoSP requires only 0.905M additional trainable parameters for its lightweight affinity head beyond the frozen depth foundation model, highlighting the efficiency of the proposed geometry-aware scale propagation.

Table 3: Ablation Study on our GeoSP Components and Designs.

(a) Geometric relations

| | Cues | | | NYU | | KITTI | |
|---|---|---|---|---|---|---|---|
| | 3D | Depth | position | Rel↓ | Ov.↓ | Rel↓ | Ov.↓ |
| Full | ✓ | ✓ | ✓ | **1.27** | **31.50** | **3.14** | **288.60** |
| A0 | | ✓ | ✓ | 1.33 | 32.10 | 3.28 | 295.14 |
| A1 | | | ✓ | 1.35 | 32.16 | 3.36 | 299.17 |

(b) Training supervision

| | Loss | | NYU | | KITTI | |
|---|---|---|---|---|---|---|
| | $\mathcal{L}_{3D}$ | $\mathcal{L}_{low}$ | Rel↓ | Ov.↓ | Rel↓ | Ov. ↓ |
| Full | ✓ | ✓ | **1.27** | **31.50** | **3.14** | **288.60** |
| B0 | | ✓ | 1.32 | 32.00 | 3.24 | 292.30 |
| B1 | ✓ | | 1.33 | 32.10 | 3.24 | 292.30 |

(c) Affinity prediction

| | Factors | | | | NYU | | KITTI | |
|---|---|---|---|---|---|---|---|---|
| | RGB | $g$ | $\phi$ | Cat. | Rel↓ | Ov. ↓ | Rel↓ | Ov. ↓ |
| Full | ✓ | ✓ | ✓ | ✓ | **1.27** | **31.50** | **3.14** | **288.60** |
| C0 | | ✓ | ✓ | ✓ | 1.40 | 31.60 | 3.80 | 291.50 |
| C1 | ✓ | | ✓ | ✓ | 1.31 | 32.00 | 3.23 | 291.90 |
| C2 | ✓ | ✓ | | ✓ | 1.32 | 32.00 | 3.25 | 292.80 |

(d) Affinity-weighted solver

| | Transfer | | NYU | | KITTI | |
|---|---|---|---|---|---|---|
| | Depth | Point | Rel↓ | Ov. ↓ | Rel↓ | Ov. ↓ |
| Full | ✓ | | **1.27** | **31.50** | **3.14** | **288.60** |
| D0 | | ✓ | 1.54 | 34.50 | 5.60 | 383.60 |

### 4.3 Ablation Study

Because the three modules of GeoSP are sequentially dependent, we ablate key designs within each stage rather than removing entire modules. We evaluate on the NYU official test set with random sampling and the KITTI Eigen test set with simulated 16-line LiDAR. Table 3 reports Rel and Ov. for four groups of ablations.

**Ablation on geometric relations.** We first examine the geometric relations introduced in Sec. 3.2. A0 removes the explicit 3D features while keeping depth-related cues and 2D position, while A1 further removes the depth cues and keeps only 2D position. Both variants degrade Rel and Ov. on NYU and KITTI. The better performance of A0 over A1 shows that depth cues provide useful information beyond 2D position. More importantly, the further gain from the full model shows that explicit 3D geometry provides stronger guidance for metric scale propagation.

**Ablation on training supervision.** We study the contribution of different supervision terms in Table 3b. B0 removes the 3D loss $\mathcal{L}3\mathrm{D}$, while B1 removes the low-frequency loss $\mathcal{L}\mathrm{low}$. Both lead to small performance drops, increasing NYU Rel from $1.27$ to $1.32/1.33$ and KITTI Rel from $3.14$ to $3.24$. These results indicate that $\mathcal{L}3\mathrm{D}$ and $\mathcal{L}\mathrm{low}$ provide complementary supervision for learning reliable metric scale propagation.

**Ablation on geometric affinity prediction.** We study the three inputs to the affinity head in Sec. 3.3 in Table 3c. Removing RGB in C0 causes a clear performance drop, while removing either $g$ or $\phi$ in C1/C2 changes Rel by at most $0.03$. This is likely because geometry has already been incorporated in neighborhood selection and $g$ and $\phi$ are partially redundant, whereas RGB provides complementary cues for affinity estimation under 3D ground-truth supervision.

**Ablation on affinity-weighted solver.** We study how metric scale is recovered by the affinity-weighted solver in Table 3d. D0 replaces the depth ratio scale constraint $s_i$ in Eq. equation 5 by adding the relative 3D displacement from $i$ to $q$ to the metric point at $i$ and projecting it onto the query ray. This increases NYU Rel from $1.27$ to $1.54$ and Ov. from $31.5$ to $34.5$. On KITTI, Rel increases from $3.14$ to $5.60$ and Ov. from $288.6$ to $383.6$. These results show that depth-space transfer provides a more effective scale propagation solver.

## 5 Conclusion

We identify that similar 2D sparse patterns can correspond to different 3D geometries, making 2D cues insufficient for reliable metric scale propagation. We propose GeoSP, which leverages 3D geometric relations to guide scale propagation. GeoSP achieves strong zero-shot generalization across unseen datasets, sensors, sparsities, and sampling patterns, with consistent gains in depth and 3D geometry. A limitation is its limited evaluation under broader sensor configurations and extremely sparse observations, which we leave for future work.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and refinement of the presentation of the methodology and experiments. All methodological decisions, experimental designs, implementations, and reported results were developed and verified by the authors. All AI-assisted content was carefully reviewed and revised by the authors, who take full responsibility for the final content of the paper.

## REFERENCES


Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2D-3D-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023.

Shenglun Chen, Xinzhu Ma, Hong Zhang, Haojie Li, and Zhihui Wang. Propagating sparse depth via depth foundation model for out-of-distribution depth completion, 2025.

Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *ECCV*, 2018.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, pp. 3354–3361, 2012.

Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *ICCV*, pp. 9443–9454, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pp. 9492–9502, 2024.

Youngjung Kim, Hyungjoo Jung, Dongbo Min, and Kwanghoon Sohn. Deep monocular depth estimation via integration of global and local predictions. *IEEE Transactions on Image Processing*, 27(8):4131–4144, 2018.

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 *CVPR*, 2025.

Yuankai Lin, Hua Yang, Tao Cheng, Wending Zhou, and Zhouping Yin. DySPN: Learning dynamic affinity for image-guided depth completion. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):4596–4609, 2024. doi: 10.1109/TCSVT.2023.3321668.

Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *AAAI*, 2021.

Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *ICRA*, 2018.

Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *CVPR*, pp. 9859–9869, 2024.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *ECCV*, 2020.

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 *CVPR*, 2019.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, pp. 31–42, 2014.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, pp. 746–760, 2012.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, pp. 567–576, 2015.

Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2021.

Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *CVPR*, 2024.

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. *arXiv preprint arXiv:1908.00463*, 2019.

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 *ICCV*, pp. 5359–5370, 2025.

Haotian Wang, Meng Yang, and Nanning Zheng. G2-MonoDepth: A general framework of generalized depth inference from monocular RGB+X data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3753–3771, 2024a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianing Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024b.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. In *ICLR*, 2026.

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.

Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *ECCV*, 2022.

Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion, 2024.

Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, pp. 899–908, 2019.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, pp. 10371–10381, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024b.

Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *CVPR*, 2023.

Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing*, 30:5264–5276, 2021.

Zhiyuan Zhou, Ruofeng Liu, Taichi Liu, Weijian Zuo, Shanshan Wang, Zhiqing Hong, and Desheng Zhang. Any to full: Prompting depth anything for depth completion in one stage, 2026.