

MetricAnything: Scaling Metric Depth Pretraining with Noisy Heterogeneous Sources

Jiahui Yang^{1,2,3}^{*}, Donglin Di^{1,2}[†], Xuancheng Zhang¹, Lei Fan⁴,
Jianxun Cui^{2,3}, Hao Li¹, and Baorui Ma¹^{*,†}

¹ Li Auto Inc., Beijing, China

² Harbin Institute of Technology, Harbin, China

³ Chongqing Research Institute of Harbin Institute of Technology, Chongqing, China

⁴ University of New South Wales, Sydney, Australia

^{*}Equal contribution. [†]Corresponding author. [‡]Project lead.

Abstract. Scaling has driven major advances in vision foundation models, yet extending it to metric depth estimation remains challenging due to heterogeneous sensor noise, camera-dependent biases, and metric ambiguity in noisy cross-source 3D data. We introduce **MetricAnything**, a simple and scalable pretraining framework that learns metric depth from diverse, noisy 3D sources without manual prompts, camera-specific modeling, or task-specific architectures. At its core is the **Sparse Metric Prompt**, formed by randomly masking depth maps, which acts as a universal interface to decouple spatial reasoning from sensor and camera biases. Pretrained on ~ 20 M image–depth pairs from reconstructed, captured, and rendered 3D data across 10,000+ camera models, MetricAnything demonstrates—for the first time—a clear scaling trend in metric depth. The model excels at prompt-driven tasks such as depth completion, super-resolution, and Radar-camera fusion, while its distilled prompt-free student achieves state-of-the-art results in monocular depth estimation, camera intrinsics recovery, single/multi-view metric 3D reconstruction, and VLA planning. Using its pretrained ViT as a visual encoder also significantly enhances Multimodal Large Language Models in spatial intelligence. These results show that metric depth estimation can follow the same scaling laws as modern foundation models, paving a scalable and efficient path toward real-world metric perception. Project page: <https://metric-anything.github.io/metric-anything-io/>

1 Introduction

Vision foundation models have achieved remarkable progress by scaling model capacity and leveraging increasingly large and diverse datasets, enabling emergent capabilities and robust generalization [22, 47, 69, 77, 79]. As AI systems increasingly interact with the physical world through robotics, AR, and autonomous driving, foundation models must go beyond 2D perception to perceive the 3D world. Monocular depth estimation (MDE) serves as a fundamental bridge, providing depth cues essential for physical interaction.

While relative depth estimation has demonstrated successful scaling through synthetic data and large-scale pseudo-label distillation [122, 123], **metric depth**

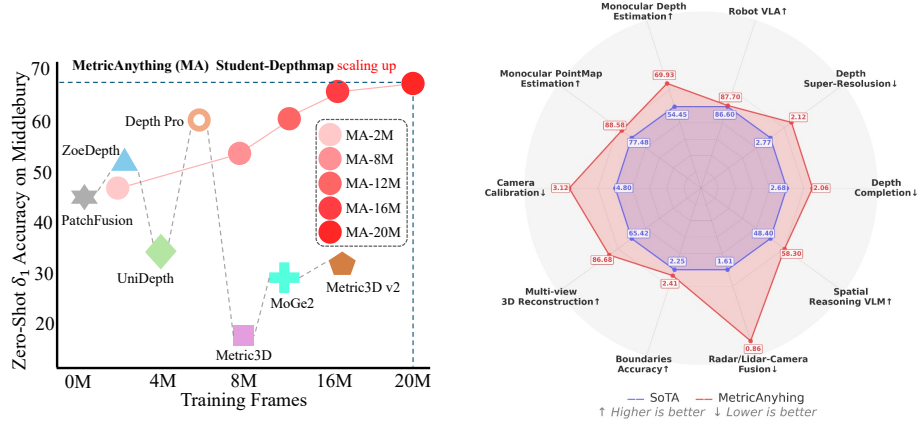


Fig. 1: Scaling and Generalization. MetricAnything exhibits a clear scaling trend and strong overall downstream performance.

estimation has not exhibited similar scaling trends in previous works, as illustrated in Fig. 1a. Unlike relative depth’s ordinal relationships, metric depth requires learning absolute, physically meaningful distances—a fundamentally harder problem compounded by data-related challenges.

The core bottleneck lies in the heterogeneity of metric depth data sources. Real-world 3D annotations come from diverse sensors (LiDAR, RGB-D, stereo cameras) or reconstruction algorithms (SfM, MVS, SLAM), each introducing distinct noise patterns, hardware-specific artifacts, and camera-dependent biases. This heterogeneity creates three critical issues: (i) significant domain gaps between sources, (ii) noisy supervision from sensor misalignment and algorithmic matching failures, and (iii) metric ambiguity due to varying camera intrinsics. Together, these factors prevent the “data soup” scaling strategy that has proven successful in other domains, limiting current metric depth methods to careful curation of small, clean datasets.

Recent approaches address these challenges with prompt-based methods [60, 64, 104, 112, 137], using sparse depth points or simulated LiDAR cues to guide prediction. However, they remain limited in scope, typically focusing on specific prompt-driven tasks with small datasets and relying on complex, hand-crafted prompt pipelines that introduce strong human priors. This task-specific engineering limits scalability and generalization.

In this work, we present a fundamentally different perspective: rather than engineering better prompts for specific tasks, we demonstrate that a simple, scalable prompt-based pretraining paradigm can unlock the metric depth perception capabilities with large-scale heterogeneous 3D data. In other words, we are not aiming to construct a stronger model for a particular prompt-driven task (e.g., depth completion or super-resolution). We posit a simple yet effective pre-training paradigm to demonstrate a new direction, which proves that with effective mitigation of data scarcity issues, scaling trends can similarly

occur in the metric estimation track, just as in NLP and 2D vision. Our key insight is that sparse prompts can serve as a universal interface to decouple spatial understanding from sensor-specific biases, enabling effective learning from diverse, noisy sources without complex engineering.

To realize this vision, we introduce **MetricAnything**, a minimalist pre-training paradigm tailored for metric depth estimation, which generates **Sparse Metric Prompts** by randomly masking portions of depth maps. This design intentionally avoids task-specific architectures, complex prompt construction rules, or strong 3D inductive biases, making specialized network design unnecessary. Instead, it enables the model to naturally learn structural and metric understanding directly from data. To support scaling, we assemble approximately 20M image–depth pairs across three categories: reconstructed 3D data (SfM/SLAM/MVS), captured 3D data (LiDAR/ToF/RGB-D), and rendered 3D data. All samples are aligned with metric annotations; for raw point clouds, we obtain per-pixel depths by projection using known calibration. The collection spans over 10,000 camera models and diverse environments, enabling unified pretraining across heterogeneous 3D sources.

The pretrained model, with its simple architecture and pretraining objective, excels in prompt-based tasks such as depth completion, super-resolution, and radar-camera depth estimation. Leveraging a dedicated distillation process, we further create a prompt-free student model that achieves state-of-the-art performance across diverse downstream tasks, including monocular depth estimation, camera calibration, multi-view metric 3D reconstruction, and Vision-Language-Action (VLA) planning. We further demonstrate that leveraging the pretrained ViT from our MetricAnything as a visual encoder substantially enhances the spatial reasoning capabilities of Multimodal Large Language Models. Remarkably, without any task-specific design, both the pretrained and distilled models consistently achieve state-of-the-art performance across these tasks.

Our findings align with “The Bitter Lesson” [96]: general-purpose, data-driven methods outperform hand-crafted designs. By showing consistent gains with scale (Fig. 1a) and robust generalization (Fig. 1b) across ten downstream tasks, we establish prompt-based pretraining as a scalable path toward general metric depth perception. Our contributions are:

- **MetricAnything:** A minimalist prompt-based pretraining paradigm that employs Sparse Metric Prompt to decouple spatial understanding from sensor biases, enabling scalable learning from heterogeneous 3D sources.
- **Demonstrated Scaling:** Aggregating 20M diverse image–depth pairs reveals stable scaling trends in metric depth, previously unseen in this domain.
- **Universal Generalization:** Pretrained and distilled models achieve state-of-the-art results across ten downstream tasks without adaptation.

2 Related Work

2.1 Monocular Depth Estimation

Relative depth estimation. Relative depth has become the dominant formulation due to its ability to learn from heterogeneous supervision, including

pairwise ordering [15, 16], stereo/MVS cues [31, 57], and pseudo-depth [82, 122]. Early CNN-based methods framed depth as regression [23, 24, 50], followed by discretization and classification strategies to address ambiguity and non-uniform distributions [6, 7, 26], though often tied to dataset-specific depth ranges. Large-scale multi-dataset training with scale-and-shift-invariant objectives proved critical: MegaDepth [57] and MiDaS [82] demonstrated the benefits of heterogeneous data mixing, while transformer architectures introduced global context beyond CNN locality [22, 81]. Depth Anything further validated scaling for relative depth through synthetic supervision and massive pseudo-label distillation (62M images), achieving strong generalization [122, 123]. Extending this success to metric depth remains challenging, as prior works [8, 10, 38, 74, 75, 108, 125] highlight the greater data demands and complexity of metric-aware perception. However, heterogeneous and noisy metric depth data limit straightforward scaling. Unlike prior approaches, our prompt-based pretraining mitigates these issues and enables scalable metric depth learning.

Metric Depth Estimation. Metric depth estimation has evolved from fixed-intrinsics RGB-D/LiDAR settings [24, 34, 124] to open-domain, camera-aware formulations that couple depth with intrinsics or canonical camera transforms [8, 10, 35, 36, 38, 73–75, 108, 125]. Representative strategies include domain-specific metric heads (ZoeDepth) [8], canonical camera normalization for zero-shot transfer (Metric3D) [38, 125], high-resolution ViTs predicting canonical inverse depth (DepthPro) [10], and direct regression of metric 3D points with camera conditioning (UniDepth) [75]. MoGe-2 decouples metric scale from affine-invariant point maps [107, 108] but relies on complex non-learnable refinement pipelines. In contrast, we avoid hand-crafted processing and propose a scalable, data-driven pretraining paradigm that learns robust metric representations from noisy, heterogeneous 3D sources.

2.2 Metric Depth Estimation with Sparse Priors

This line of work predicts dense metric depth from RGB with sparse cues such as LiDAR, RGB-D, or radar [18, 19, 70, 98, 130, 136]. Representative methods include Marigold-DC [104], OMNI-DC [137], PriorDA [112], PromptDA [60], and TacoDepth [111], which leverage diffusion guidance, gradient integration, metric-prior alignment, prompt injection, or sensor-specific fusion. Despite progress, most approaches remain limited to specific prompt types or sensor modalities, often relying on heavy hand-designed prompt simulation and small-scale datasets. In contrast, we minimize human prior design and instead scale simple pretext tasks over naturally heterogeneous data. This enables robust spatial understanding and state-of-the-art performance across multi-sensor fusion, monocular depth estimation, camera calibration, multi-view 3D reconstruction, and VLA planning without task-specific specialization.

3 Method

We aim to develop a general-purpose MDE foundation model that achieves both fine structural fidelity and accurate metric perception, enabling robust general-

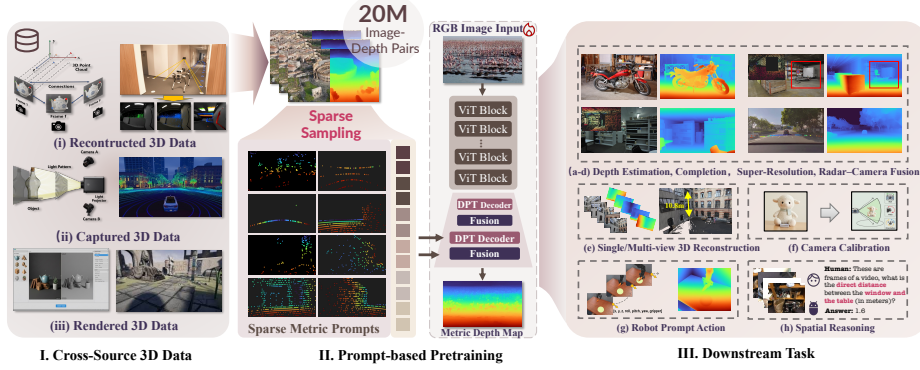


Fig. 2: Overview of MetricAnything. (I) We aggregate diverse open-source 3D data into per-pixel metric depth maps, forming a $\sim 20\text{M}$ image-depth dataset captured by over 10,000 cameras across heterogeneous scenes. (II) Sparse Metric Prompts, generated by randomly masking depth maps, provide a minimal interface that decouples spatial reasoning from sensor and camera biases, enabling metric depth learning from noisy, heterogeneous sources. (III) The pretrained model and its distilled prompt-free student generalize robustly across multiple downstream tasks, revealing a clear scaling trend and establishing a solid foundation for versatile, data-driven metric perception.

ization across diverse tasks. As shown in Fig. 2, we aggregate a 20M-scale dataset (Sec. 3.1) to support a sparse metric prompt pretraining paradigm that handles data heterogeneity (Sec. 3.2). Finally, we distill the learned knowledge into a prompt-free student model (Sec. 3.3).

3.1 Multi-Source Data Collection

To enable scalable pretraining, we aggregate open-source 3D datasets with metric annotations and standardize all inputs as metric depth maps with validity masks:

$$\{\mathbf{G}, \mathbf{M}\}, \quad \mathbf{G} \in \mathbb{R}^{H \times W}, \quad \mathbf{M} \in \{0, 1\}^{H \times W}, \quad (1)$$

where $\mathbf{G}$ denotes the metric depth along the camera z -axis, and $\mathbf{M}$ indicates valid measurements ($\mathbf{M}(p) = 1$ for valid pixels and $\mathbf{M}(p) = 0$ otherwise).

Reconstructed 3D Data. We aggregate depth maps from open-source datasets [4, 5, 20, 40, 101, 105, 113, 119] derived from diverse reconstruction techniques, including SfM, SLAM, MVS, and binocular/multi-view stereo (*e.g.*, plane-sweeping, PatchMatch). While these maps provide per-pixel metric depth $\mathbf{G}$ with validity masks $\mathbf{M}$, artifacts remain prevalent in challenging regions—such as textureless areas, specular surfaces, thin structures, occlusions, and dynamic scenes—primarily due to correspondence matching failures.

Captured 3D Data. We also incorporate point clouds captured by LiDAR, ToF, and RGB-D sensors [1, 14, 21, 29, 33, 37, 78, 93, 95, 114, 117, 127]. Given the camera intrinsics $\mathbf{K}$ and sensor-to-camera pose $\mathbf{T}_{c \leftarrow s} = [\mathbf{R} \mid \mathbf{t}]$, depth maps $\{\mathbf{G}, \mathbf{M}\}$ are obtained by projecting each sensor-frame point $\mathbf{X}_j^s$ onto the corresponding camera plane. Sensor-derived depth is often noisy and sparse due

to limitations in sensor resolution and the influence of environmental factors such as severe weather. Additional sources of error include beam divergence and sampling gaps (affecting thin structures), weak or anomalous returns on specular, metallic, or transparent/semi-transparent surfaces, multi-path interference, mixed pixels at sharp boundaries, out-of-range distances, occlusions, and motion-induced time offsets. These factors result in invalid measurements, indicated by $\mathbf{M}(p) = 0$ for affected pixels.

Rendered 3D Data. To complement real-world data, we incorporate synthetic depth maps from virtual engines [2, 25, 27, 32, 32, 56, 83, 85, 106, 110, 115, 133]. Despite limited realism, these noise-free maps preserve fine structural details and sharp geometry, providing valuable geometric supervision.

In total, this purposefully aggregated dataset comprises approximately 20 million image–depth pairs, captured by over 10,000 distinct camera models and spanning a wide variety of real and synthetic scenes.

3.2 Pretraining via Sparse Metric Prompt

Problem Formulation. Given a monocular image $I \in \mathbb{R}^{H \times W \times 3}$ and a corresponding sparse metric prompt:

$$P = \{(x_i, y_i, d_i)\}_{i=1}^N, \quad (2)$$

where each triplet (x_i, y_i, d_i) represents a pixel coordinate (x_i, y_i) and its associated metric depth value d_i . The objective of pretraining is to learn a function f_θ , parameterized by θ , that predicts a dense metric depth map D from the input image I , conditioned on the prompt P :

$$D = f_\theta(I, P), \quad D \in \mathbb{R}^{H \times W}. \quad (3)$$

Here, prompt P provides metric depth for a small subset of pixels, offering limited geometric cues. The model f_θ learns to propagate these sparse metric constraints across the image, yielding a complete, geometrically consistent, and metrically accurate depth map. Since our data sources inevitably contain inherent noise and varying patterns of incompleteness, f_θ is trained to identify and correct potential errors within the sparse prompt P .

Prompt Preparation. From each source in $\{\mathbf{G}, \mathbf{M}\}$, we randomly sample $N \in [2,000, 40,000]$ valid pixels (around 1% of all image pixels) from its depth map to construct a sparse metric prompt P and its corresponding mask M' . Due to diverse patterns, sparsity, and incompleteness across $\{\mathbf{G}\}$, the sparsely sampled prompt P inevitably inherits heterogeneous distributions and irregular structures, which challenge models to effectively handle diverse distributions.

To mitigate this, we first apply a prompt preparation step that maps all prompts into a shared intermediate domain. Following PriorDA [112], we apply a pre-trained depth prediction model [10] to obtain a prior depth map P_d . Then, inspired by See3D [67], we perform Pixel-wise Depth Scale Alignment (PDSA) and Global Metric Depth Recovery (GMDR) to fill the missing regions in P under the guidance of P_d . The final input prompt is constructed by concatenating the PDSA-refined map, the GMDR-corrected map, and the prompt mask M' along

the channel dimension. This yields a unified and regularized representation of shape $H \times W \times 3$, avoiding issues with irregular prompt network design and training efficiency, while also easing the handling of diverse prompts.

Prompt Injection. Although this preparation process can maintain metric consistency between P_d and Prompt P through pixel-wise and global alignment, its parameter-free (non-learning) nature remains sensitive to sampling patterns and noise propagated from prompt P . To further address this, we employ a prompt-injection mechanism that allows the model to correct noisy prompts and generate accurate, dense depth predictions. Following prior works, several strategies can be used for prompt injection, including adaptive layer normalization (AdaLN) [71], cross-attention [84, 103], ControlNet-style conditioning [128], conditioned DPT heads [60], and conditioning layers parallel to the RGB input [112]. Among these, we opt for a conditioned DPT head, as it introduces only about 5% extra parameters while maintaining efficiency through lightweight interpolation and shallow convolutional layers.

This design allocates most parameters to the foundational depth backbone, allowing the main network pathway to correct noise and structural inconsistencies, rather than relying on the injected condition layers. Importantly, we make no bespoke modifications to the backbone; instead, we reuse a general, efficient, and widely validated architecture [10]. The only difference is that we merge the patch encoder and image encoder into a single shared ViT to achieve a balance between accuracy and efficiency, as detailed in our ablation studies. This emphasizes our deliberate shift away from hand-crafted priors and manually designed prompts, relying entirely on data-driven learning. By isolating conditioning from architectural changes, we can cleanly assess the effectiveness of the learned prompts and rigorously study the scaling behavior.

Pretraining Objective. Following DepthPro [10], we use mean absolute error (MAE) and scale-and-shift invariant mean absolute gradient error (SSI-MAGE) losses for synthetic data. For real-world data, we adopt a robust MAE loss that drops top- n -largest-loss regions per image during training to mitigate noise sensitivity, where n is set as 20%. The overall objective is:

$$L_{total} = \alpha L_{MAE} + \beta L_{SSI-MAGE}. \quad (4)$$

3.3 Prompt-Free Model Distillation

Distillation Process. To validate the benefits of large-scale pretraining and transfer them to scenarios where prompts are unavailable, we distill the pre-trained (teacher) model into a prompt-free student that performs dense depth prediction solely from RGB inputs. Specifically, we leverage the trained teacher model to generate high-fidelity pseudo depth maps conditioned on sparse prompt P for all real-world images in the collected dataset via a simple feed-forward process, effectively transforming sparse and noisy prompts into dense, high-quality supervision signals (i.e., pseudo labels). Unlike conventional RGB-D or LiDAR sensors—whose effective sensing range is constrained by hardware to only a few to tens of meters—and many depth-estimation algorithms that struggle to estimate distant backgrounds, our teacher-generated pseudo labels span diverse

environments and cover both near and far distances. These teacher labels train a prompt-free monocular depth estimation model, preserving the teacher’s metric understanding without prompts.

Student Selection. A straightforward approach is to directly reuse the pre-training network architecture as the student backbone while removing the prompt layers. However, in our preliminary experiments, this approach resulted in suboptimal performance when trained

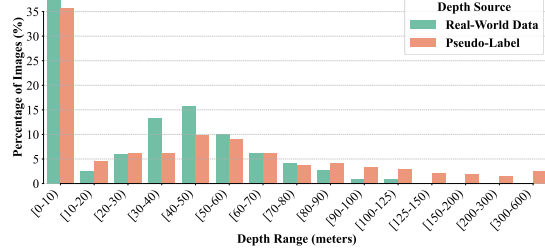


Fig. 3: Percentile depth range comparison from seven datasets (real-world vs. our pseudo labels).

on the teacher labels, despite many prior works [8, 10, 38, 108] successfully training similar ViT-DPT configurations from scratch. We attribute this suboptimality to two main factors. First, unlike most real-world datasets biased toward short-range depths, our teacher model generates reliable pseudo labels across both near and far regions (Fig. 3). This wider depth distribution exposes the limitations of conventional depth losses: direct depth supervision (L1/L2) blurs fine geometric details, whereas inverse-depth loss decays too rapidly with distance, losing effective supervision for distant regions. Consequently, these standard losses are suboptimal for supervising our high-fidelity, large-range pseudo labels. Second, typical ViT encoders with DPT decoders employ U-Net-style skip connections to propagate shallow features upward. This design stabilizes training under noisy supervision because low-level cues (*e.g.*, textures and colors) are inherently more consistent and easier to learn. By routing these low-level features directly to the output head, the network mitigates conflicts with noisy targets, thereby smoothing gradient fluctuations. However, this heavily relies on low-level details, underutilizing the high-level semantic cues essential for precise depth reasoning. In our setting, the pseudo labels are generated by a unified model, exhibit minimal domain gaps, and have been refined to eliminate most noise. Thus, we can reduce the reliance on shallow-to-deep feature injection and explore more advanced network designs that fully exploit the rich semantic representations from the deep blocks of the ViT encoder.

Improved Student. Building upon the above observations, we retain the multi-scale fusion mechanism of [10] while introducing two modifications. First, we adopt a distance-balanced log-depth loss to improve supervision across varying depth ranges. Specifically, instead of directly regressing metric depth, we transform the predicted and ground-truth depths into log space to balance gradients between near and far regions. This preserves fine-grained sensitivity in close-range areas while preventing long-range depths from being under-supervised. Given a metric depth value x (in meters), its log-transformed representation is:

$$D_{\log} = 1 - \frac{\ln(x)}{\ln(C)}, \quad (5)$$

where C is a hyperparameter that controls the trade-off between long-range and short-range supervision.

Second, leveraging the high-fidelity teacher labels, we invert the conventional skip-connection scheme between the ViT encoder and DPT head: injecting deep, high-level ViT features into the deeper decoding layers near the output, while shallow, low-level features feed into the shallower decoder layers, as shown in Fig. 4. This inversion

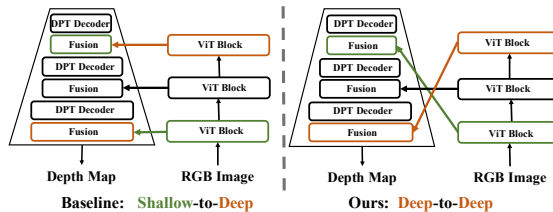


Fig. 4: Skip-Connection in ViT-DPT Architecture.

emphasizes semantic reasoning at the final prediction stage, fully exploiting the rich semantic cues embedded in the teacher-generated pseudo labels, which already exhibit low noise and high structural consistency. In this way, we experimentally demonstrate that the student model achieves stable, prompt-free metric depth estimation, effectively distilling the metric perception capability from the pre-trained model.

4 Experiment

4.1 Implementation Details

We train the pre-trained model on 144 H200 GPUs for 100k steps with a 10k-step warm-up, using peak learning rates of 1×10^{-6} for the ViT backbone and 1×10^{-5} for the DPT head and prompt layer. The additional convolutional kernels and biases required for prompt injection are zero-initialized. For the distilled prompt-free student, pseudo-labels are generated by the pre-trained model. We evaluate two training paradigms. (1) *From-scratch training*: the ViT backbone is initialized from DINOv3 ViT-H+/16 [90], with a randomly initialized DPT head. Training runs on 144 H200 GPUs for 200k steps with a 5k-step warm-up and peak learning rates of 2×10^{-6} (ViT) and 2×10^{-5} (DPT). (2) *Fine-tuning*: all parameters are initialized from MoGe-2 [108] and trained on 80 H200 GPUs for 10k steps using a 1k-step warm-up, with peak learning rates of 2.5×10^{-6} (ViT) and 2.5×10^{-5} (DPT). Additional training details are provided in the *Supp.*

4.2 Prompt-Based Downstream Task

Zero-Shot Depth Super-Resolution and Completion. We study depth super-resolution and completion in a zero-shot setting, where our pretrained model directly takes sparse or low-resolution depth maps as prompts, without any task-specific finetuning. We compare two categories of baselines: (a) post-aligned MDE, including DepthAnything V2 [123] and DepthPro [10]; and (b) prior-based MDE, including LingBot-Depth [97], PromptDA [60], PriorDA [112], DepthLab [64], Omni-DC [137], and Marigold-DC [104]. Following PriorDA [112],

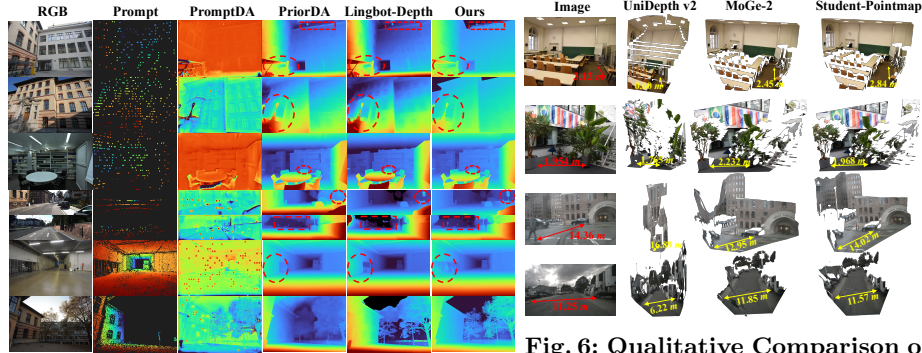


Fig. 5: Visualization of Depth SR and Point Maps. The **red arrows** indicate **Completion**. Our method better recovers GT distance, the **yellow** indicate pre-dicted distance.

Fig. 6: Qualitative Comparison of

Table 1: Zero-shot Depth Super-Resolution / Completion (AbsRel %).

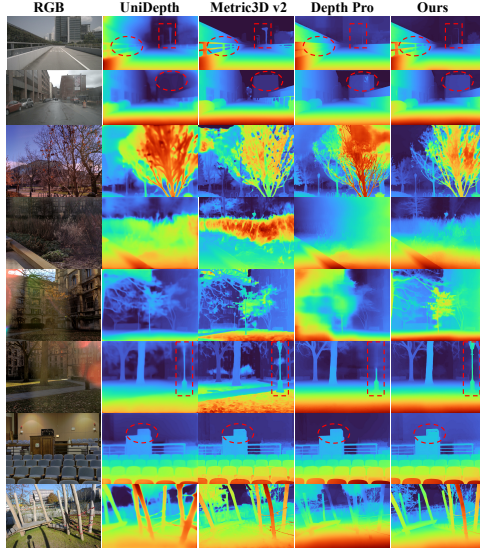
Method	NYUv2 ↓				ETH-3D ↓				KITTI ↓				Avg. Rank
	8×	16×	LiDAR	Extreme	8×	16×	LiDAR	Extreme	8×	16×	LiDAR	Extreme	
DAv2 [123]	4.77	5.13	4.85	4.77	6.27	7.38	7.41	6.51	9.54	11.22	8.86	9.25	7.92
DepthPro [10]	4.48	4.83	4.47	4.41	5.88	6.79	5.31	6.51	6.76	9.16	6.05	6.19	6.58
Omni-DC [137]	1.57	3.11	2.12	2.63	1.86	4.09	1.88	1.98	4.05	8.35	5.27	4.17	3.75
Marigold-DC	1.83	3.32	1.90	2.13	2.33	4.75	2.27	2.03	5.17	9.47	6.88	5.62	5.17
DepthLab [64]	2.60	3.73	4.30	6.30	2.60	4.50	6.40	8.01	17.17	22.90	37.17	40.29	7.58
PromptDA [60]	1.61	1.75	17.59	16.96	1.80	2.56	18.86	18.18	3.92	4.95	21.96	21.39	5.50
PriorDA [112]	1.73	2.79	2.01	2.01	2.06	3.91	1.90	1.61	4.54	8.20	4.81	3.76	3.25
LingBot-Depth [97]	2.23	3.05	2.19	2.03	1.18	1.54	13.19	4.12	4.24	7.13	6.55	3.26	3.75
Ours-Pretrain	1.53	1.86	1.70	2.08	1.46	2.03	0.87	0.84	2.34	3.53	3.53	3.36	1.50

we construct four prompt types—LiDAR-like sparse scans, extremely sparse samples (100 points), and depth maps downsampled by 8× and 16×. All prompts are directly fed into our pretrained model for zero-shot inference on unseen datasets, including NYUv2 [89], ETH3D [87], and KITTI [30], covering indoor, outdoor, and open-world scenarios. As shown in Tab. 1 and Fig. 5, our pretrained model demonstrates strong zero-shot performance across all prompt types and datasets, consistently outperforming both post-aligned and prior-based baselines. Unlike previous approaches [112,137] that generate synthetic prompts (*e.g.*, LiDAR simulation or noisy downsampling) to approximate test-time conditions, our model is trained only once with simple, sparsely sampled prompts and operates fully zero-shot manner, without any task-specific design or prompt alignment.

Radar-Camera Depth Estimation. We evaluate the transferability of our pretrained model to an unseen sensing modality on the nuScenes dataset [12]. Radar signals are extremely sparse and noisy, exhibiting a substantial modality gap from the sparse metric prompts used during pretraining. Notably, no radar data is included in our pretraining corpus. Following the standard protocol, we conduct two settings: (1) training the teacher model from scratch using radar signals as prompts and LiDAR depth as supervision; and (2) finetuning our pretrained model under the same supervision. As shown in Tab. 2, finetuning achieves state-of-the-art performance, nearly doubling the accuracy of its

Table 2: Radar-Camera Depth Estimation on nuScenes [12] (mm).

Type	Method	0-50m		0-70m		0-80m	
		MAE $\downarrow$	RMSE $\downarrow$	MAE $\downarrow$	RMSE $\downarrow$	MAE $\downarrow$	RMSE $\downarrow$
Independent	Lin <i>et al.</i> [62]	2034.9	4316.5	2294.7	5338.2	2371.0	5623.0
	RC-PDA [66]	2225.0	4156.5	3326.1	6700.6	3713.6	7692.8
	RDyn [28]					6434.0	
	DORN [65]	1926.6	4124.8	2380.6	5252.7	2467.7	5554.3
	Singh <i>et al.</i> [91]	1727.7	3746.8	2073.2	4500.7	2179.3	4898.7
	CaFNet [94]	1674.2	3674.5	2010.3	4493.1	2109.8	4765.6
	Li <i>et al.</i> [55]	1524.5	3567.3	1822.9	4303.6	1927.0	4609.6
	TacoDepth [111]	1423.6	3275.8	1712.6	3960.5	1833.4	4150.2
Plug-in	RadarCam-Depth [54]	1286.1	2964.3	1587.9	3662.5	1689.7	3948.0
	TacoDepth [111]	1046.8	2487.5	1347.1	3152.8	1492.4	3324.8
MetricAnything-Pretrain	From-Scratch	1335.4	2958.8	1622.9	3788.1	2101.7	4033.2
	Finetune	651.4	2084.4	863.6	2771.6	934.5	3057.5

**Fig. 7: Zero-shot Visual Comparison.** Student-DepthMap excels in preserving fine-grained details and handling complex illumination, surpassing competing methods.

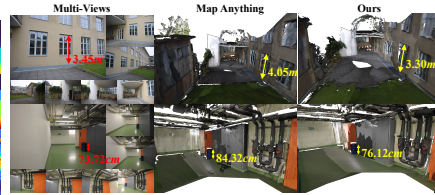
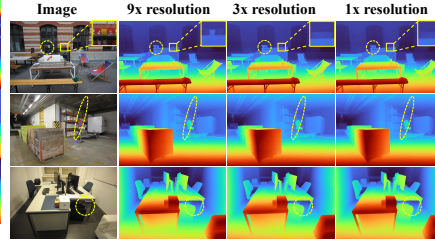
from-scratch counterpart and outperforming existing radar-image fusion methods. These results indicate that our pretraining paradigm learns versatile metric representations that generalize effectively to new sensing modalities.

4.3 Prompt-Free Downstream Tasks

Monocular Depth Estimation. We evaluate prompt-free metric prediction in two settings: monocular depth map estimation and monocular geometry estimation. For depth map estimation, we train a dedicated student network from scratch (Sec. 3.3), denoted as **Student-DepthMap**. We compare against recent methods including Depth Anything [122], Depth Anything V2 [123], Metric3D [125], Metric3D V2 [38], PatchFusion [58], UniDepth [75], ZoeDepth [8], and DepthPro [10]. Following standard practice, we report δ_1 on six benchmarks—Booster [80], Middlebury [86], Sun-RGBD [92], ETH3D [88], NuScenes [12], and Sintel [11]—under a strict zero-shot protocol. Quantitative and qualitative

Table 3: Monocular depth estimation performance of our *Student-DepthMap* against prior methods, as evidenced by higher δ_1 accuracy (% , $\uparrow$).

Method	ETH3D	Booster	NuScenes	SunRGBD	Sintel	Middlebury	Rank $\downarrow$
DepthAnything [122]	9.3	52.3	35.4	85.0	6.9	39.3	6.00
DAV2 [123]	36.3	59.5	17.7	72.4	5.9	37.2	6.17
Metric3D [125]	34.2	4.7	64.4	16.9	17.3	13.6	6.83
Metric3D-v2 [38]	87.7	39.4	82.6	75.6	38.3	29.9	4.17
PatchFusion [58]	51.8	22.6	20.4	53.6	14.0	49.9	6.00
UniDepth [75]	25.3	27.6	83.6	95.8	16.5	31.9	5.00
ZoeDepth [8]	34.2	21.6	28.1	85.7	7.8	53.8	5.83
DepthPro [10]	41.5	46.6	49.1	89.0	40.0	60.5	3.17
Student-DepthMap	79.9	59.5	88.1	97.7	27.7	65.8	1.50

**Fig. 8:** Auxiliary monocular depth inputs improve performance of MapAnything. Red: GT distance; Yellow: reconstruction distance.**Fig. 9: Test-time Resolution Scaling.** Qualitative results at 1 $\times$, 3 $\times$, and 9 $\times$ resolution. Higher resolutions recover finer details.

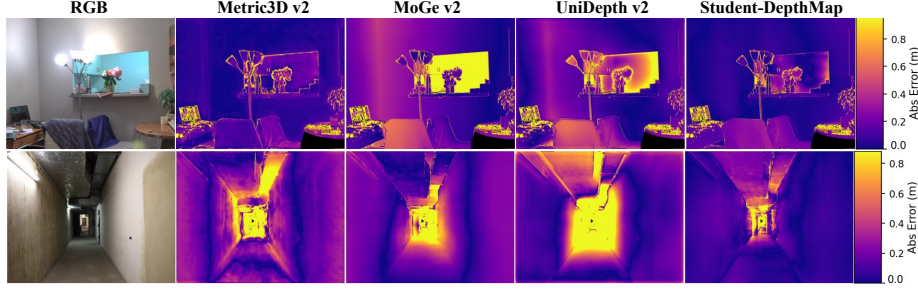


Fig. 10: Zero-shot visual comparison of absolute depth error maps.

Table 4: Monocular Geometry Estimation without Camera Intrinsics.

Method	KITTI [30]		ETH3D [88]		iBims-1 [48]		DIODE [102]	
	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$
ZoeDepth [8]	17.0	85.4	57.1	33.7	17.4	67.2	39.3	29.3
MAS3R [52]	56.7	9.84	47.2	20.1	18.7	61.5	54.9	19.0
DAV1 [122]	11.6	94.5	40.2	24.0	12.9	81.8	58.0	16.2
DAV2 [123]	10.6	88.6	36.1	36.3	11.1	91.7	41.2	22.1
UniDepth V1 [75]	4.69	98.4	56.9	14.9	23.8	57.6	17.1	71.9
UniDepth V2 [74]	8.58	95.4	20.7	69.5	9.52	93.2	43.0	51.8
Depth Pro [10]	23.5	38.3	38.5	32.8	15.9	81.5	31.9	37.7
MoGe-2 [108]	18.1	62.9	10.4	90.8	13.6	83.0	17.5	66.4
DAV3-Metric-Large [59]	11.8	83.0	11.0	90.1	27.6	43.5	24.2	53.7
DAV3-Nested-G-L [59]	83.2	0.00	80.7	0.00	39.1	20.9	51.9	26.4
Student-PointMap	3.22	97.7	10.8	90.2	11.3	86.6	13.9	79.8

Table 5: Monocular Geometry Estimation with Provided Camera Intrinsics.

Our *Student-PointMap* model outperforms baselines, demonstrating superior robustness under known camera parameters.

Method	KITTI [30]		ETH3D [88]		iBims-1 [48]		DIODE [102]	
	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$	AbsRel $\downarrow$	$\delta_1 \uparrow$
Metric3D V2 [38]	5.25	98.0	11.8	88.8	9.96	94.1	49.1	1.98
UniDepth V1 [75]	4.43	98.5	44.5	26.7	22.6	60.5	21.0	63.5
UniDepth V2 [74]	5.98	97.7	15.0	85.2	7.71	95.5	41.0	67.1
MoGe-2 [108]	8.64	93.7	10.5	92.2	9.92	92.4	16.2	77.1
Student-PointMap	5.47	99.6	9.75	92.4	9.47	92.0	14.1	74.3

Table 6: Evaluation results for camera calibration (in degrees).

Method	ETH3D [88]		iBims-1 [48]		Average		
	Mean $\downarrow$	Med. $\downarrow$	Mean $\downarrow$	Med. $\downarrow$	Mean $\downarrow$	Med. $\downarrow$	Rank $\downarrow$
Perspective [43]	13.6	11.9	10.6	9.30	12.1	10.6	7.25
WildCam [135]	7.70	5.81	9.48	9.08	8.59	7.45	5.00
Depth Pro [10]	7.18	6.34	4.12	2.58	5.65	4.46	3.50
LeReS [126]	8.26	7.19	18.4	17.5	13.33	12.35	7.00
DUST3R [109]	5.77	3.60	3.83	2.53	4.80	3.07	2.25
UniDepth [75]	10.7	9.96	11.9	5.96	11.3	7.96	6.25
MoGe-2 [108]	4.66	3.04	8.17	6.77	6.42	4.91	3.25
Student-PointMap	2.50	1.86	3.74	3.24	3.12	2.55	1.50

Table 7: LIBERO benchmark success rates by task category.

Method	Spatial	Object	Goal	Long	Avg
TraceVLA [134]	84.6	85.2	75.1	54.1	74.8
Octo-Base [100]	78.9	85.7	84.6	51.1	75.1
OpenVLA [46]	84.7	88.4	79.2	53.7	76.5
SpatialVLA [76]	88.2	89.9	78.6	55.5	78.1
CoT-VLA [132]	87.5	91.6	87.6	69.0	83.9
NORA-AC [41]	85.6	89.4	80.0	63.0	79.5
WorldVLA [13]	87.6	96.2	83.4	60.0	79.1
π_0 -FAST [72]	96.4	96.8	88.6	60.2	85.5
ThinkAct [39]	88.3	91.4	87.1	70.9	84.4
Baseline-DAV2 [123]	87.0	95.4	87.6	77.2	86.6
Ours	88.6	94.4	88.8	78.8	87.7

results are shown in Tab. 3, Fig. 7 and Fig. 10, with additional metrics (AbsRel [49], Log_{10} , δ_2/δ_3 , RMSE) provided in the *Supp.* For monocular geometry estimation, we fine-tune state-of-the-art model (*i.e.*, MoGe-2 [108]) using pseudo-labels generated by our pretrained model, denoted as **Student-PointMap** (in Fig. 6, Tab. 5 and Tab. 4). This setting evaluates the generality and precision of our pseudo-label distillation across different architectures and prediction heads. Consistent gains are observed under both training-from-scratch and fine-tuning regimes. More visualizations and analyses are provided in the *Supp.*

Recovering Camera Intrinsics. We further show that the predicted point maps from **Student-PointMap** enable accurate recovery of camera intrinsics via focal length optimization. Under mild assumptions (unit aspect ratio and centered principal point), our method achieves the best mean and median angular errors compared with point-map-based optimization approaches [75, 109, 126] and learning-based calibration methods [43, 135] (Tab. 6).

VLA Planning. We extend MetricAnything to VLA frameworks for embodied planning, achieving SOTA performance (Tab. 7). Following MolmoACT [51], we

Table 8: Results on VSI-Bench [121]. We use 16 input frames for Spatial-MLLM/Qwen2.5VL [3] and 1 FPS for Gemini-1.5 Pro [99], while others follow VSI-Bench defaults (8-32 frames).

Methods	Numerical Question				Multi-Choice Question				Avg. Rank	
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order		
Proprietary Models										
GPT-4o [42]	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0	8
Gemini-1.5 Pro [99]	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6	45.4	3
Open-source Models										
InternVL2-40B [17]	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6	36.0	7
LongVILA-8B [118]	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5	21.6	13
VILA-1.5-40B [61]	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2	10
LongVA-7B [129]	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2	12
LLaVA-OneVision-72B [53]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2	5
LLaVA-Video-72B [131]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9	4
Qwen2.5VL-3B [3]	24.3	24.7	31.7	22.6	38.3	41.6	26.3	21.2	30.6	11
Qwen2.5VL-7B [3]	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8	33.0	9
Qwen2.5VL-72B [3]	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9	37.0	6
Spatial-MLLM-4B [116]	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3	48.4	2
Ours	70.0	51.5	67.5	65.0	66.2	76.8	40.2	29.0	58.3	1

Table 10: Performance gains from integrating our Student module into MapAnything.

Method	ETH3D [88]		ScanNet [21]		Avg	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
MapAnything [45]	20.43	69.07	37.61	61.78	29.02	65.42
MapAnything + Ours	18.98	73.94	5.90	99.41	12.44	86.68

Table 9: Zero-shot Depth Boundary (F1 score [10]) of Student-DepthMap.

Method	Sintel F1↑	Spring F1↑	iBims F1↑	Rank↓
<i>Relative</i>				
DepthAnything [122]	0.261	0.0109	0.127	7.33
DepthAnything v2 [123]	0.228	0.0610	0.111	8.00
Marigold [44]	0.068	-	0.149	7.50
<i>Absolute</i>				
DPT [81]	0.181	-	0.113	9.00
Metric3D [125]	0.037	-	0.055	11.50
Metric3D v2 [38]	0.321	0.0723	0.096	6.33
ZoeDepth [8]	0.027	0.0043	0.035	11.33
PatchFusion [58]	0.312	-	0.134	6.00
UniDepth [75]	0.316	0.0017	0.039	8.67
Depth Pro [10]	0.409	0.1100	0.176	2.00
UniDepth-v2 [74]	0.344	0.0737	0.138	4.00
MoGe-v2 [108]	0.282	0.0890	0.194	3.67
Student-DepthMap	0.382	0.1635	0.179	1.67

Table 11: Single-frame inference latency under FP32 (ms).

Method	Param	Res.	FLOPs	Latency (ms)		
				VGA	HD	4K
DepthPro [10]	952M	1.5k	8.8T	246.0	246.0	246.0
Ours-PreTrain	993M	1.5k	20.0T	285.2	285.2	285.3
Student-PointMap	326M	840	4.0T	29.4	369.8	5043.9
Student-DepthMap	877M	1.5k	21.5T	278.8	278.5	278.8

distill our prompt-free model by supervising depth token prediction rather than using depth as input. Benchmarking against Depth Anything V2 [123], we report success rates on four tasks from LIBERO [63].

Spatial Understanding. We integrate MetricAnything’s ViT encoder into a VLM as a 3D-aware backbone, fusing 3D and 2D tokens via cross-attention to inject metric priors. Built on LLaVA-Next-Video-7B [9] and fine-tuned on 200K QA pairs and 4,225 planning instances, our model leads VSI-Bench [120] (58.3 avg., Tab. 8). This validates that our transferable 3D priors significantly enhance spatial reasoning. See *Supp* for more details..

Zero-shot Boundaries Accuracy. We evaluate the sharpness of the predicted geometry by our ‘Student-DepthMap’ using two synthetic benchmarks, Spring [68] and Sintel [11], together with the real-world iBims-1 [48] dataset. As reported in Tab. 9, our method achieves the best average accuracy boundaris compared to DepthPro [10] and MoGe-2 [108].

Multi-view Metric 3D Reconstruction. Integrating our model into MapAnything [45] as a plug-and-play depth module boosts performance on ETH3D [88] and ScanNet [21] (Tab. 10). This implies better cross-view consistency. Notably, Fig. 8 highlights superior metric-scale recovery, demonstrating that our strong priors effectively transfer to multi-view tasks.

Runtime. We benchmark inference latency at VGA (640×480), HD (1920 × 1080), and 4K (4032 × 3024) resolutions under FP32 on an H200 GPU, reporting runtime with and without prompts, as well as parameter counts and FLOPs (Tab. 11). Leveraging a shared ViT encoder, our pretrain-model incurs negligible latency overhead compared to Depth Pro, despite the additional prompt branch.

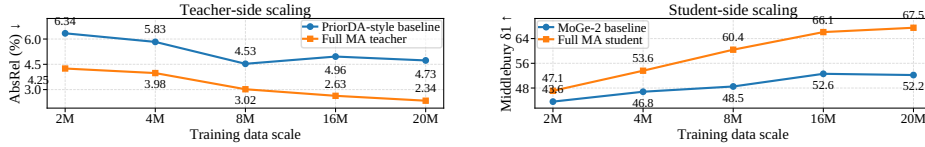


Fig. 11: Baseline data scaling comparison. We compare Full MA with a PriorDA-style teacher baseline and a MoGe-2 student baseline under different data scales. Full MA exhibits more favorable scaling behavior on both the teacher side and the student side, while the controlled baselines show weaker or saturated gains at larger scales.

Table 12: Component-wise scaling attribution. AbsRel% ($\downarrow$).

Setting	SMP	Hetero.	Inv.	D-B	Booster	ETH3D	Avg.
Full MA	✓	✓	✓	✓	28.2	14.7	21.5
Specific-prompt MA	✗	✓	✓	✓	34.1 (+17.3)	33.2 (+55.7)	33.7 (+36.2)
Single-source MA	✓	✗	✓	✓	45.4 (+37.9)	17.4 (+15.5)	31.4 (+31.5)
U-Net-style student	✓	✓	✗	✓	39.7 (+29.0)	32.0 (+54.1)	35.9 (+40.1)
w/o dist.-balanced loss	✓	✓	✓	✗	40.3 (+30.0)	16.1 (+8.7)	28.2 (+23.8)

For the prompt-free variants (Student-DepthMap/Student-PointMap), inference time scales with model capacity.

4.4 Ablation Study

Scaling up Data. We pretrain our model on data subsets ranging from 5% to 100%. On KITTI with $8\times$ super-resolution, AbsRel steadily decreases from 5.22 to 2.34, indicating improved generalization as the data scale increases (Fig. 1a). To further validate this scaling behavior, we build two controlled baselines. On the teacher side, we follow the prompt processing and loss design of PriorDA and train a PriorDA-style baseline under the same data scales, 100k training steps, and evaluation protocol. From 8M to 20M samples, full MetricAnything (MA) reduces AbsRel by 22.5%, whereas the PriorDA-style baseline degrades by 4.4%. On the student side, we distill the same student architecture using MoGe-2-generated pseudo labels to isolate the effect of the teacher-label source (Fig. 11). From 16M to 20M, the MoGe-2 baseline slightly drops by 0.8%, while full MA further improves by 2.1% and maintains a clear accuracy advantage.

Scaling Attribution. We perform component-wise attribution under the same student data, training schedule, and evaluation setup. The green numbers in Tab. 12 indicate the relative AbsRel reduction of Full MetricAnything over each ablated variant. We ablate four components: replacing random sparse-mask prompts with PriorDA-style hand-designed prompts, using only single-source synthetic/rendered data, replacing the inverse-skip student with U-Net-style fusion, and removing the distance-balanced distillation loss. All ablations degrade performance, showing that the student scaling gain relies on the joint design of Sparse Metric Prompt pretraining, heterogeneous data, inverse-skip distillation, and distance-balanced supervision, rather than data scale alone.

Network Architecture. We benchmark our inverse skip design against U-Net [81] using both real and pseudo-labeled data. As shown in Tab. 13, while U-Net works well on real data, it underutilizes pseudo labels. In contrast, our

Table 13: Network architecture ablation on NuScenes and ETH3D.

Method	NuScenes [12]				ETH3D [88]			
	Real-world		Pseudo		Real-world		Pseudo	
	Abs↓	RMSE↓	Abs↓	RMSE↓	Abs↓	RMSE↓	Abs↓	RMSE↓
UNet-style	0.213	9.631	0.187	7.222	0.327	5.353	0.269	1.440
Ours	0.235	9.662	0.125	6.267	0.334	5.490	0.182	1.898

Table 14: Ablation on Objectives across Depth Ranges.

Training objective	DIODE [102] AbsRel ↓					
	0-10m	10-20m	10-30m	30-40m	40-50m	>50m
Depth [38]	0.574	0.582	0.599	0.622	0.645	0.689
Log-depth [8]	0.556	0.562	0.577	0.594	0.612	0.663
Inverse-depth [10]	0.467	0.493	0.565	0.581	0.592	0.632
Ours	0.465	0.480	0.489	0.502	0.537	0.589

Table 15: Two Ablation Studies: (left) number of prompt points on Hypersim; (right) balance coefficient C on DIODE (AbsRel ↓).

Points	500	1000	2000	4000	8000	16000	32000	64000	C	50	100	200	400	600
AbsRel↓	0.043	0.041	0.038	0.036	0.034	0.033	0.032	0.031	0-40 m	0.441	0.452	0.454	0.456	0.465
Time (ms)	224	243	256	266	274	284	299	308	>40 m	0.635	0.611	0.565	0.562	0.559

design effectively leverages distillation. Notably, our pseudo labels consistently boost performance across all architectures, validating their quality.

Training Objectives. We compare standard losses (depth, log, inverse) against our proposed distance-balanced loss (Eq. 5). As shown in Tab. 14, our method matches inverse-depth at close range while significantly outperforming it at long distances. This validates the effectiveness of our loss for the broad depth ranges inherent in our pseudo-labels (Fig. 3).

Prompt Setting. We analyze the impact of prompt density by varying sampled pixels N from 500 to 64,000. As shown in Tab. 15 (left), accuracy gains on Hypersim saturate as N increases. Consequently, we select $N \in [2,000, 40,000]$ to balance accuracy and efficiency. Notably, thanks to data scaling, our model generalizes robustly to extremely sparse prompts (*e.g.*, $N = 100$ in Tab. 1).

Balance Weights. We investigate the balance weight C (Eq. 5) by evaluating candidates from 50 to 600 (Tab. 15, right). Visualizations in the **Supp** reveal that excessive weights (*e.g.*, 600) bias the model towards distant regions at the cost of near-field details, while insufficient weights compromise distant supervision. Consequently, we select $C = 400$ for the optimal trade-off.

Test-Time Resolution Scaling. As shown in Fig. 9, our model naturally supports test-time resolution scaling. When increasing the input resolution from $1\times$ to $3\times$ and $9\times$, the predicted depth maps recover finer structures and sharper object boundaries, without any additional fine-tuning.

Supp Overview. (a) Generalizability evaluations on unseen sensors, scenarios, and extreme conditions; (b) Extended discussions on additional experiments and ablation studies; (c) More implementation details.

5 Conclusion

We present **MetricAnything**, a scalable framework learning from heterogeneous sources without manual prompts. Utilizing Sparse Metric Prompts to decouple geometry from sensor priors, we reveal the first clear scaling law in metric depth. Our models achieve SOTA results across downstream tasks, offering a robust solution for real-world depth perception.

References

1. Antequera, M.L., Gargallo, P., Hofinger, M., Buló, S.R., Kuang, Y., Kotschieder, P.: Mapillary planet-scale depth dataset. In: European Conference on Computer Vision. pp. 589–604. Springer (2020)
2. Apollo, B.: Apollo synthetic dataset (2019), accessed: 2025-6-12
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025)
4. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., 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. Bauer, Z., Gomez-Donoso, F., Cruz, E., Orts-Escolano, S., Cazorla, M.: Uasol, a large-scale high-resolution outdoor stereo dataset. Scientific data **6**(1), 162 (2019)
6. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4009–4018 (2021)
7. Bhat, S.F., Alhashim, I., Wonka, P.: Localbins: Improving depth estimation by learning local distributions. In: European Conference on Computer Vision. pp. 480–496. Springer (2022)
8. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023)
9. Bhat, V., Lan, Y.H., Krishnamurthy, P., Karri, R., Khorrami, F.: 3d cavla: Leveraging depth and 3d context to generalize vision language action models for unseen tasks. arXiv preprint arXiv:2505.05800 (2025)
10. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations (2025)
11. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)
12. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
13. Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al.: Worldvla: Towards autoregressive action world model. arXiv preprint arXiv:2506.21539 (2025)
14. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017)
15. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. Advances in neural information processing systems **29** (2016)
16. Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., Deng, J.: Oasis: A large-scale dataset for single image 3d in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
17. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning

- for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
18. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: AAAI (2020)
 19. Cheng, X., Wang, P., Yang, R.: Learning depth with convolutional spatial propagation network. TPAMI (2019)
 20. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
 21. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scan-net: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
 22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
 23. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
 24. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* **27** (2014)
 25. Fonder, M., Van Droogenbroeck, M.: 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. pp. 0–0 (2019)
 26. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2002–2011 (2018)
 27. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016)
 28. Gasperini, S., Koch, P., Dallabetta, V., Navab, N., Busam, B., Tombari, F.: R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In: 2021 International Conference on 3D Vision (3DV). pp. 751–760. IEEE (2021)
 29. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters* **6**(3), 4947–4954 (2021)
 30. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving. The KITTI vision benchmark suite. In: CVPR **2**, 5 (2012)
 31. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3828–3838 (2019)
 32. Gómez, J.L., Silva, M., Seoane, A., Borrás, A., Noriega, M., Ros, G., Iglesias-Guitián, J.A., López, A.M.: All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *Neurocomputing* **637**, 130038 (2025)

33. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2485–2494 (2020)
34. Guizilini, V., Vasiljevic, I., Chen, D., Ambrus, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
35. Guizilini, V., Vasiljevic, I., Chen, D., Ambrus, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
36. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth any camera: Zero-shot metric depth estimation from any camera. In: CVPR (2025)
37. Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V., Ondruska, P.: One thousand and one hours: Self-driving motion prediction dataset. arXiv preprint arXiv:2006.14480 (2020)
38. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: 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)
39. Huang, C.P., Wu, Y.H., Chen, M.H., Wang, Y.C.F., Yang, F.E.: Thinkact: Vision-language-action reasoning via reinforced visual latent planning. arXiv preprint arXiv:2507.16815 (2025)
40. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2821–2830 (2018)
41. Hung, C.Y., Sun, Q., Hong, P., Zadeh, A., Li, C., Tan, U., Majumder, N., Poria, S., et al.: Nora: A small open-sourced generalist vision language action model for embodied tasks. arXiv preprint arXiv:2504.19854 (2025)
42. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
43. Jin, L., Zhang, J., Hold-Geoffroy, Y., Wang, O., Blackburn-Matzen, K., Sticha, M., Fouhey, D.F.: Perspective fields for single image camera calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17307–17316 (2023)
44. Ke, B., Obukhov, A., Huang, S., Metzger, N., Dautt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
45. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kotschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction (2025), arXiv preprint arXiv:2509.13414
46. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)
47. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

48. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. pp. 0–0 (2018)
49. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 89–96 (2014)
50. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: *2016 Fourth international conference on 3D vision (3DV)*. pp. 239–248. IEEE (2016)
51. Lee, J., Duan, J., Fang, H., Deng, Y., Liu, S., Li, B., Fang, B., Zhang, J., Wang, Y.R., Lee, S., et al.: Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917* (2025)
52. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: *European Conference on Computer Vision*. pp. 71–91. Springer (2024)
53. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024)
54. Li, H., Ma, Y., Gu, Y., Hu, K., Liu, Y., Zuo, X.: Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 10665–10672. IEEE (2024)
55. Li, H., Jing, M., Jin, W., Dong, S., Liang, J., Fan, H., Ji, R.: Sparse beats dense: Rethinking supervision in radar-camera depth completion. In: *European Conference on Computer Vision*. pp. 127–143. Springer (2024)
56. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3205–3215 (2023)
57. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2041–2050 (2018)
58. Li, Z., Bhat, S.F., Wonka, P.: Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10016–10025 (2024)
59. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025)
60. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: *CVPR* (2025)
61. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoenybi, M., Han, S.: Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 26679–26689 (2023)
62. Lin, J.T., Dai, D., Van Gool, L.: Depth estimation from monocular images and sparse radar data. in *2020 IEEE International Conference on Intelligent Robots and Systems (IROS)*. pp. 10233–10240 (2020)
63. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems* **36**, 44776–44791 (2023)

64. Liu, Z., Cheng, K.L., Wang, Q., Wang, S., Ouyang, H., Tan, B., Zhu, K., Shen, Y., Chen, Q., Luo, P.: Depthlab: From partial to complete. *arXiv preprint arXiv:2412.18153* (2024)
65. Lo, C.C., Vandewalle, P.: Depth estimation from monocular images and sparse radar using deep ordinal regression network. In: *2021 IEEE International Conference on Image Processing (ICIP)*. pp. 3343–3347. IEEE (2021)
66. Long, Y., Morris, D., Liu, X., Castro, M., Chakravarty, P., Narayanan, P.: Radar-camera pixel depth association for depth completion. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 12507–12516 (2021)
67. Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., Wang, X.: You see it, you got it: Learning 3d creation on pose-free videos at scale. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 2016–2029 (2025)
68. Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4981–4991 (2023)
69. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023)
70. Park, J.H., Jeong, C., Lee, J., Jeon, H.G.: Depth prompting for sensor-agnostic depth estimation. In: *CVPR* (2024)
71. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4195–4205 (2023)
72. Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., Levine, S.: Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747* (2025)
73. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: UniK3D: Universal camera monocular 3d estimation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025)
74. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110* (2025)
75. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10106–10116 (2024)
76. Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al.: Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830* (2025)
77. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
78. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., et al.: Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238* (2021)

79. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
80. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Booster: a benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(1), 85–102 (2023)
81. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
82. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: 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 (2020)
83. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
84. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
85. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: 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. pp. 3234–3243 (2016)
86. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: German conference on pattern recognition. pp. 31–42. Springer (2014)
87. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
88. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)
89. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
90. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khali-dov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
91. Singh, A.D., Ba, Y., Sarker, A., Zhang, H., Kadambi, A., Soatto, S., Srivastava, M., Wong, A.: Depth estimation from camera image and mmwave radar point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9275–9285 (2023)
92. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)

93. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019)
94. Sun, H., Feng, H., Ott, J., Servadei, L., Wille, R.: Cafnet: A confidence-driven framework for radar camera depth estimation. In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 2734–2740. IEEE (2024)
95. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2446–2454 (2020)
96. Sutton, R.: The bitter lesson. *Incomplete Ideas (blog)* **13**(1), 38 (2019)
97. Tan, B., Sun, C., Qin, X., Adai, H., Fu, Z., Zhou, T., Zhang, H., Xu, Y., Zhu, X., Shen, Y., Xue, N.: Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895* (2026)
98. Tang, J., Tian, F.P., An, B., Li, J., Tan, P.: Bilateral propagation network for depth completion. In: *CVPR* (2024)
99. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024)
100. Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al.: Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213* (2024)
101. Tosi, F., Liao, Y., Schmitt, C., Geiger, A.: Smd-nets: Stereo mixture density networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8942–8952 (2021)
102. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463* (2019)
103. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
104. Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K., Obukhov, A.: Marigold-dc: Zero-shot monocular depth completion with guided diffusion (2024)
105. Wang, K., Shen, S.: Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters* **5**(2), 3307–3314 (2020)
106. Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., Chu, X.: Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678* (2019)
107. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5261–5271 (2025)
108. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details (2025)
109. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 20697–20709 (2024)

110. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916. IEEE (2020)
111. Wang, Y., Li, J., Hong, C., Li, R., Sun, L., Song, X., Wang, Z., Cao, Z., Lin, G.: Tacodepth: Towards efficient radar-camera depth estimation with one-stage fusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10523–10533 (2025)
112. Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H., Zhao, Z.: Depth anything with any prior (2025)
113. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundation-stereo: Zero-shot stereo matching. CVPR (2025)
114. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., et al.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493 (2023)
115. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705 (2018)
116. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. ArXiv **abs/2505.23747** (2025)
117. Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., et al.: Pandaset: Advanced sensor suite dataset for autonomous driving. In: 2021 IEEE international intelligent transportation systems conference (ITSC). pp. 3095–3101. IEEE (2021)
118. Xue, F., Chen, Y., Li, D., Hu, Q., Zhu, L., Li, X., Fang, Y., Tang, H., Yang, S., Liu, Z., He, E., Yin, H., Molchanov, P., Kautz, J., Fan, L., Zhu, Y., Lu, Y., Han, S.: Longvila: Scaling long-context visual language models for long videos. ArXiv **abs/2408.10188** (2024)
119. Yang, G., Song, X., Huang, C., Deng, Z., Shi, J., Zhou, B.: Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 899–908 (2019)
120. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10632–10643 (2025)
121. Yang, J., Yang, S., Gupta, A.W., Han, R., Li, F.F., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. ArXiv **abs/2412.14171** (2024)
122. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
123. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024)
124. Yin, W., Liu, Y., Shen, C., Yan, Y.: Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5684–5693 (2019)
125. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9043–9053 (2023)

126. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 204–213 (2021)
127. Zamir, A.R., Sax, A., Shen, W., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3712–3722 (2018)
128. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
129. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. *ArXiv abs/2406.16852* (2024)
130. Zhang, Y., Guo, X., Poggi, M., Zhu, Z., Huang, G., Mattoccia, S.: Completion-former: Depth completion with convolutions and vision transformers. In: *CVPR* (2023)
131. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. *ArXiv abs/2410.02713* (2024)
132. Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al.: Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 1702–1713 (2025)
133. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: *European Conference on Computer Vision*. pp. 519–535. Springer (2020)
134. Zheng, R., Liang, Y., Huang, S., Gao, J., Daumé III, H., Kolobov, A., Huang, F., Yang, J.: Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345* (2024)
135. Zhu, S., Kumar, A., Hu, M., Liu, X.: Tame a wild camera: In-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems* **36**, 45137–45149 (2023)
136. Zuo, Y., Deng, J.: Ogni-dc: Robust depth completion with optimization-guided neural iterations. In: *ECCV* (2024)
137. Zuo, Y., Yang, W., Ma, Z., Deng, J.: Omni-dc: Highly robust depth completion with multiresolution depth integration. *ICCV* (2025)