MetricAnything: Scaling Metric Depth Pretraining with Noisy Heterogeneous Sources

Baorui Ma[‡], **Jiahui Yang**[*], **Donglin Di**[‡], **Xuancheng Zhang**,
Jianxun Cui, **Hao Li**, **Xie Yan**, **Wei Chen**

Li Auto Inc

Abstract

Scaling has powered recent advances in vision foundation models, yet extending this paradigm 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 **Metric Anything**, a simple and scalable pretraining framework that learns metric depth from noisy, diverse 3D sources without manually engineered prompts, camera-specific modeling, or task-specific architectures. Central to our approach is the **Sparse Metric Prompt**, created by randomly masking depth maps, which serves as a universal interface that decouples spatial reasoning from sensor and camera biases. Using ∼20M image–depth pairs spanning reconstructed, captured, and rendered 3D data across 10,000+ camera models, we demonstrate—for the first time—a clear scaling trend in the metric depth track. The pretrained 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 on monocular depth estimation, camera intrinsics recovery, single/multi-view metric 3D reconstruction, and VLA planning. We also show that using pretrained ViT of Metric Anything as a visual encoder significantly boosts Multimodal Large Language Model capabilities in spatial intelligence. These results show that metric depth estimation can benefit from the same scaling laws that drive modern foundation models, establishing a new path toward scalable and efficient real-world metric perception. We open-source MetricAnything at https://metric-anything.github.io/metric-anything-io/ to support community research.

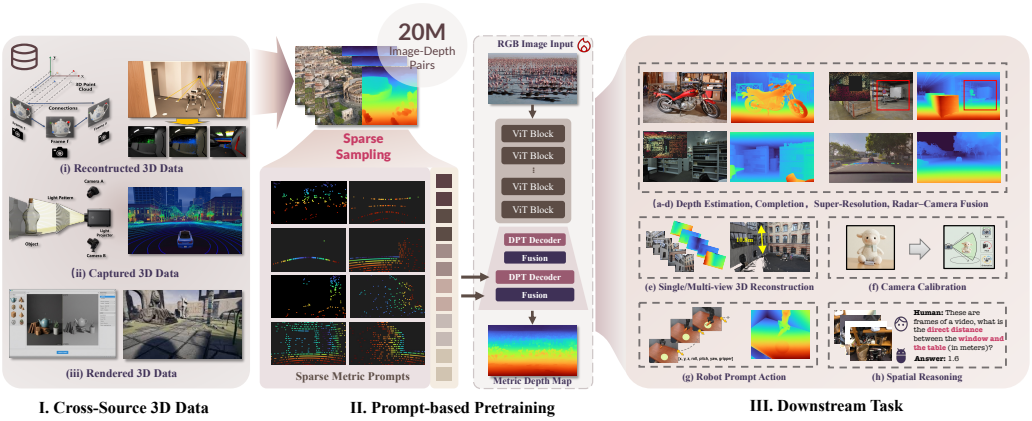


Figure 1: **Overview of Metric Anything.** (I) We aggregate diverse open-source 3D data into per-pixel metric depth maps, forming a ∼20M image–depth dataset captured by over 10,000 cameras across heterogeneous scenes. (II) Sparse Metric Prompts, generated by randomly masking depth maps, provide a minimal 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.

[*]Equal contribution.

[†]Corresponding author. Correspondence to {mabaorui2014@gmail.com}.

[‡]Project leader.

Contents

1 Introduction **3**

2 Related Work **4**

2.1 Monocular Depth Estimation 4

2.2 Metric Depth Estimation with Sparse Priors 5

3 Method **5**

3.1 Multi-Source Data Collection 5

3.2 Pretraining via Sparse Metric Prompt 6

3.3 Prompt-Free Model Distillation 7

4 Experiment **8**

4.1 Prompt-Based Downstream Task 8

4.1.1 Zero-Shot Depth Super-Resolution and Completion. 8

4.1.2 Radar-Camera Depth Estimation. 9

4.2 Prompt-Free Downstream Tasks 10

4.2.1 Monocular Depth Estimation. 10

4.2.2 Recovering Camera Intrinsics. 14

4.2.3 Zero-shot Boundaries Accuracy Measure 16

4.2.4 Multi-view Metric 3D Reconstruction. 16

4.2.5 VLA Planning 17

4.2.6 Spatial Understanding of MLLMs 18

5 Ablation Study **20**

5.1 Scaling up Data. 20

5.2 Network Architecture. 20

5.3 Runtime. 21

5.4 Test Time Resolution Scaling 21

5.5 Training Objectives 21

5.6 Prompt Setting 22

5.7 Balance Weights 23

6 Generalizability to Unseen Sensors, Scenarios, and Extreme Environmental Conditions **24**

6.1 Generalization across Sensor Configurations 24

6.2 Robustness under Environmental Degradation 24

6.3 Generalization to Unseen Visual Domains 25

7 Training Details **30**

7.1 Training and Test Set Split. 30

7.2 Training Details of Pre-trained Model. 30

7.3 Training Details of Distilled Model. 30

7.4 Training Details of Vision-Language-Model. 30

7.5 Training Details of Vision-Language-Action Model. 30

7.6 Loss Functions 32

8 Limitations **33**

9 Conclusion **33**

1 Introduction

Vision foundation models have achieved remarkable progress through scaling—training larger models on ever-expanding datasets to unlock emergent capabilities and robust generalization [23, 70, 47, 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 [124, 125], **metric depth estimation has not exhibited similar scaling trends in previous works**, as illustrated in Fig. 2a. 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 have attempted to address these challenges through prompt-based methods [60, 114, 140, 64, 106], using sparse depth points or simulated LiDAR cues to guide depth prediction. However, these methods remain limited in scope—they typically focus on specific prompt-driven downstream tasks with small datasets and rely on complex, hand-crafted prompt construction pipelines that introduce strong human priors. This task-specific engineering limits both scalability and generalization.

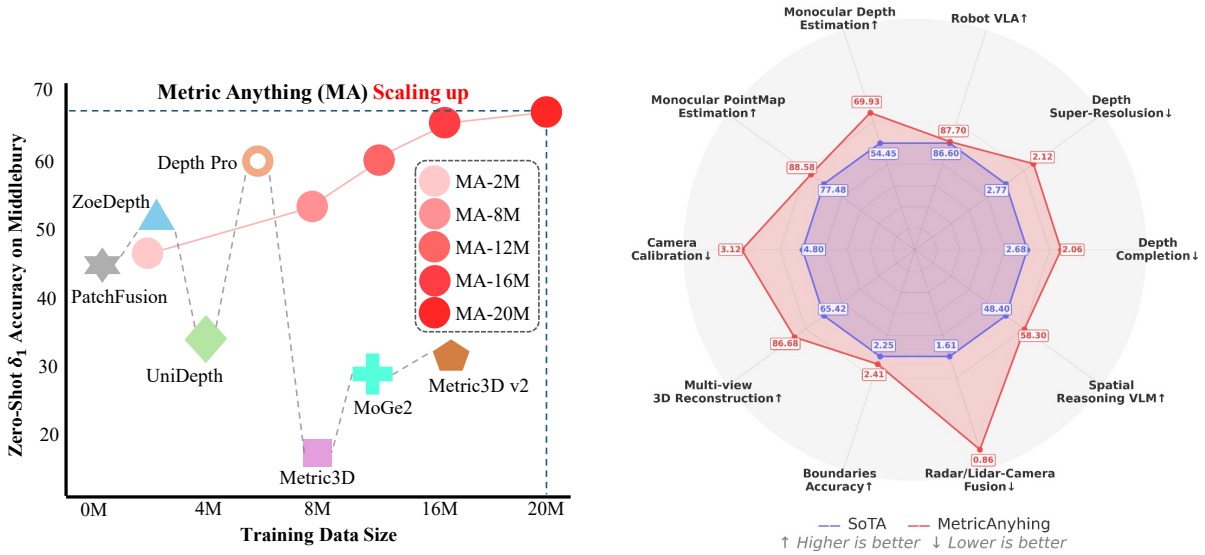
In this work, we present a fundamentally different perspective: rather than engineering better prompts for specific tasks, we demonstrate that a simple, scalable prompt-based pretraining paradigm can unlock the metric depth perception capabilities with large-scale heterogeneous 3D data. In other words, 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 **Metric Anything**, a minimalist pretraining 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 pretext 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 Metric Anything 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" [98]: general-purpose, data-driven methods systematically outperform hand-crafted designs. By demonstrating consistent performance gains with increasing scale (see Fig. 2a) and robust generalization (Fig. 2b) across ten downstream tasks, we establish prompt-based pretraining as a scalable path toward general-purpose metric depth perception. Our contributions are:

- **Metric Anything**: 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.



(a) **Scaling trend.** Larger training dataset yields consistently higher zero-shot δ_1 accuracy.

(b) **Task performance.** Radar plot over downstream tasks; Larger area indicates better performance.

Figure 2: **Scaling and Generalization.** MetricAnything exhibits a clear scaling trend and strong overall downstream performance.

- **Demonstrated Scaling**: Aggregation of 20M diverse depth-image pairs reveals stable scaling trends in metric depth estimation, previously unseen in this domain.
- **Universal Generalization**: Both pretrained and distilled models achieve state-of-the-art performance across ten downstream tasks without task-specific engineering.

2 Related Work

2.1 Monocular Depth Estimation

Relative depth estimation. Over the past decade, relative depth has emerged as the dominant formulation because it can learn from heterogeneous and weakly supervised signals—pairwise orderings [15, 16], stereo/MVS cues [32, 57], and pseudo-depth [82, 124]—yielding broad data coverage and strong cross-domain generalization. Early CNNs treated depth as regression [25, 24, 50]. Subsequent work introduced adaptive discretizations and per-pixel classification [27, 6, 7] to better handle non-uniform distributions and ambiguity, but these methods often coupled models to dataset-specific depth ranges, limiting transfer. A key advance was large-scale, multi-dataset training with scale-and-shift-invariant objectives: MegaDepth [57] highlighted the benefits of diverse training data, MiDaS [82] established mixing heterogeneous datasets with scale-and-shift-invariant losses, and transformer-based architectures brought global context beyond CNN locality constraints [23, 81]. Depth Anything further demonstrated a practical scaling route for relative depth by combining synthetic supervision that is noise-free and highly consistent with massive pseudo-labell distillation (62 million images), producing strong generalization [124, 125]. Replicating this success for metric depth remains an open challenge. Prior works [38, 127, 8, 10, 110, 75, 74] have recognized that metric-aware perception is inherently more challenging than relative depth perception and requires substantially more data for effective learning. However, metric depth data, originating from diverse algorithms or hardware configurations, exhibits various types of noise and heterogeneous patterns, further impeding the benefits obtainable from scaling. Unlike prior approaches, our prompt-based pretraining paradigm explicitly addresses this inherent noise and heterogeneity, establishing a viable scaling path for robust metric depth estimation.

Metric Depth Estimation. Metric depth estimation has progressed from domain-constrained, fixed-intrinsics settings with limited RGB-D/LiDAR supervision [25, 126, 35] to camera-aware, open-domain formulations that couple depth with intrinsics or canonical camera transforms [8, 127, 38, 75, 10, 110]. Representative strategies include attaching domain-specific metric heads to relative-depth backbones (ZoeDepth) [8], learning canonical camera normalization for zero-shot transfer (Metric3D and its scaled variants) [127, 38], predicting canonical inverse depth with high-resolution ViTs (DepthPro) [10], and directly regressing metric 3D points with camera-

conditioned features (UniDepth) [75]. MoGe-2 further decouples metric scale from affine-invariant point maps [110, 109]. While MoGe-2 seeks gains via complex, non-learnable pipelines (e.g., edge detection and depth refinement) to exploit large-scale real data. In this work, we discard hand-crafted processing and propose a pretraining paradigm from a data-driven perspective that learns a robust representation of metric depth from noisy and heterogeneous 3D sources, enabling the model to convert coarse data into a scalable source of accurate pseudo-labels.

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 [71, 132, 139, 19, 18, 100]. Representative approaches include: Marigold-DC, which repurposes a diffusion model for guided denoising with dynamic scale and shift to achieve zero-shot densification [106]; OMNI-DC, which predicts multi-scale depth gradients and integrates them to reduce long-range drift, augmented with a Laplacian loss for ambiguous regions and scale normalization for cross-dataset transfer [140]; and methods like PriorDA, which align metric priors with relative depth and refine structure under metric and geometric constraints using pixel-level alignment and distance-aware reweighting [114]. Other efforts explore foundation model prompting, such as PromptDA, which injects low-resolution noisy depth via a multi-scale fusion decoder but relies on heavy data pipelines involving LiDAR noise simulation, 3D reconstruction hole filling, and mixed supervision [60]. Sensor-specific designs have also emerged, e.g., TacoDepth, which fuses radar and camera data in a single stage using graph-based radar structure extraction and pyramid fusion [113]. Despite these advances, most methods remain confined to a single prompt type (e.g., for depth completion or super-resolution) or a single sensor modality. While PriorDA attempts to handle more diverse prompts, it still focuses on hand-designed prompt simulation—such as creating missing areas, downsampling GT, or adding outliers and blur—and remains limited to small-scale training datasets. In contrast, our work shows that minimizing human prior design in prompts and instead learning from naturally heterogeneous data offers a more general and scalable approach. Through simple pretext tasks and data scaling, our model learns essential and robust spatial understanding, achieving state-of-the-art performance across diverse downstream tasks—including multi-sensor fusion, monocular depth estimation, camera calibration, multi-view 3D reconstruction, and VLA policy—without being confined to any single sub-task.

3 Method

We aim to develop a general-purpose MDE foundation model that achieves both fine structural fidelity and accurate metric perception, enabling robust generalization across diverse tasks. To this end, we first aggregate a 20M-scale image–depth dataset (Sec. 3.1) from open-source 3D datasets. We introduce a sparse metric prompt-based pretraining paradigm that supports scalable learning from heterogeneous sources (Sec. 3.2). Finally, the pretrained knowledge is distilled into a prompt-free student model for broader applicability (Sec. 3.3). An overview is shown in Fig. 1.

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:

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

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 collect depth maps from open-source datasets [4, 40, 103, 116, 122, 5, 20, 107] produced by a wide range of reconstruction techniques, such as SfM, SLAM, MVS, binocular/multi-view stereo (e.g., plane-sweeping, PatchMatch), and dense temporal stereo. These depth maps are represented as per-pixel metric depth $\mathbf{G}$ with validity mask $\mathbf{M}$, as defined above. Such maps often contain artifacts and missing regions, especially in weakly textured areas, specular or metallic surfaces, thin structures, sharp boundaries, repetitive patterns, depth discontinuities, occlusions, or under motion blur and dynamic scenes. These errors typically arise from incorrect or lost matches during reconstruction.

Captured 3D Data. We also incorporate point clouds captured by LiDAR, ToF, and RGB-D sensors [30, 120, 37, 129, 78, 117, 14, 95, 97, 1, 34, 21]. 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

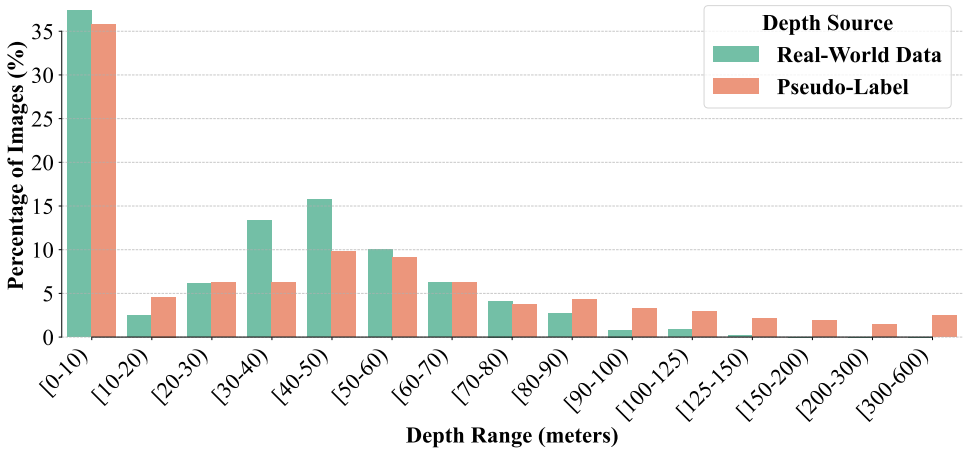


Figure 3: Percentile Depth Range Comparison from Seven Datasets (Real-world vs. Our Pseudo Labels).

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 the real-world data, we also include a small portion of depth maps rendered from virtual engines [118, 26, 84, 28, 108, 112, 2, 56, 33, 136, 33, 86]. Although these synthetic depth maps exhibit limited visual realism and diversity, they are completely noise-free, preserve fine structural details, and display sharp geometry, thereby 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,

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

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:

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

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 [114], we apply a pre-trained depth prediction model [10] to obtain a prior depth map P_d. Then, inspired by See3D [68], 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.

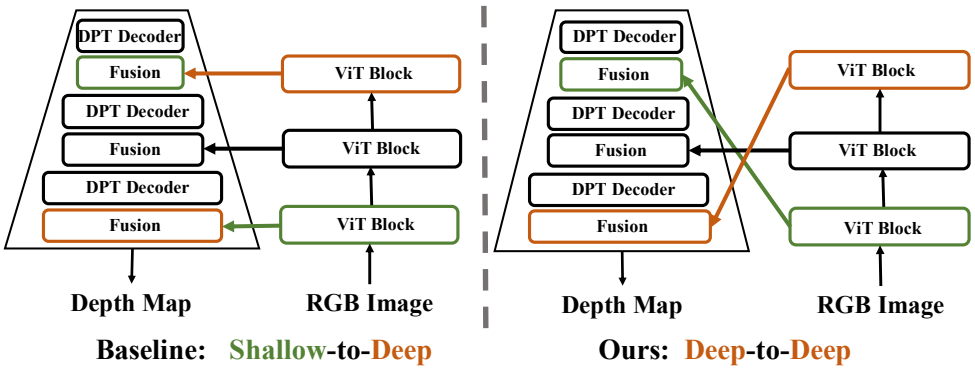


Figure 4: Skip-Connection in ViT-DPT Architecture.

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) [72], cross-attention [105, 85], ControlNet-style conditioning [130], conditioned DPT heads [60], and conditioning layers parallel to the RGB input [114]. 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.

Pre-training 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 training objective is formulated as:

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

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 on the teacher labels, despite many prior works [10, 110, 38, 8] successfully training similar ViT–DPT configurations from scratch.

We hypothesize two main causes for the suboptimality. First, our teacher model generates reliable pseudo labels across both near and far regions, unlike most real-world datasets biased toward short-range depths (see Fig. 3 for a comparison). This wider depth distribution exposes limitations of conventional depth losses: direct depth supervision (L1/L2) blurs fine geometric details, while 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-head decoders use U-Net-style skip connections, injecting shallow features into deeper layers and propagating deep features upward. While this stabilizes training under noisy supervision, since the low-level cues in the ViT encoder (e.g., textures and colors) are more consistent and easier to learn. The low-level feature is connected via a skip connection to the DPT head near the output, which mitigates conflicts between the output and the noisy supervision, thereby smoothing gradient fluctuations. However, it can underutilize high-level semantic cues essential for precise depth. In our setting, pseudo labels are generated by a unified model, exhibit minimal domain gap, and have been refined to correct most noisy. As a result, we can reduce reliance on shallow-to-deep feature injection and explore more aggressive network designs that fully exploit the rich semantic cues provided by deep block of the ViT encoder.

Improved Student. Drawing on these observations and analyses, we retain the multi-scale fusion mechanism proposed by [10] while introducing two key modifications. First, we design a distance-balanced inverse-depth loss that preserves fine-grained sensitivity in near regions while extending effective supervision to long-distance areas. Depth values in log-space are defined as:

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

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 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 Prompt-Based Downstream Task

4.1.1 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 [125] and DepthPro [10]; and (b) prior-based MDE, including LingBot-Depth [99], [60], PriorDA [114], DepthLab [64], Omni-DC [140], and Marigold-DC [106]. Following PriorDA [114], 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 [91], ETH3D [89], and KITTI [31], 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 [114, 140] 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. This enables superior generalization across diverse prompt densities, spatial layouts, and scene domains.

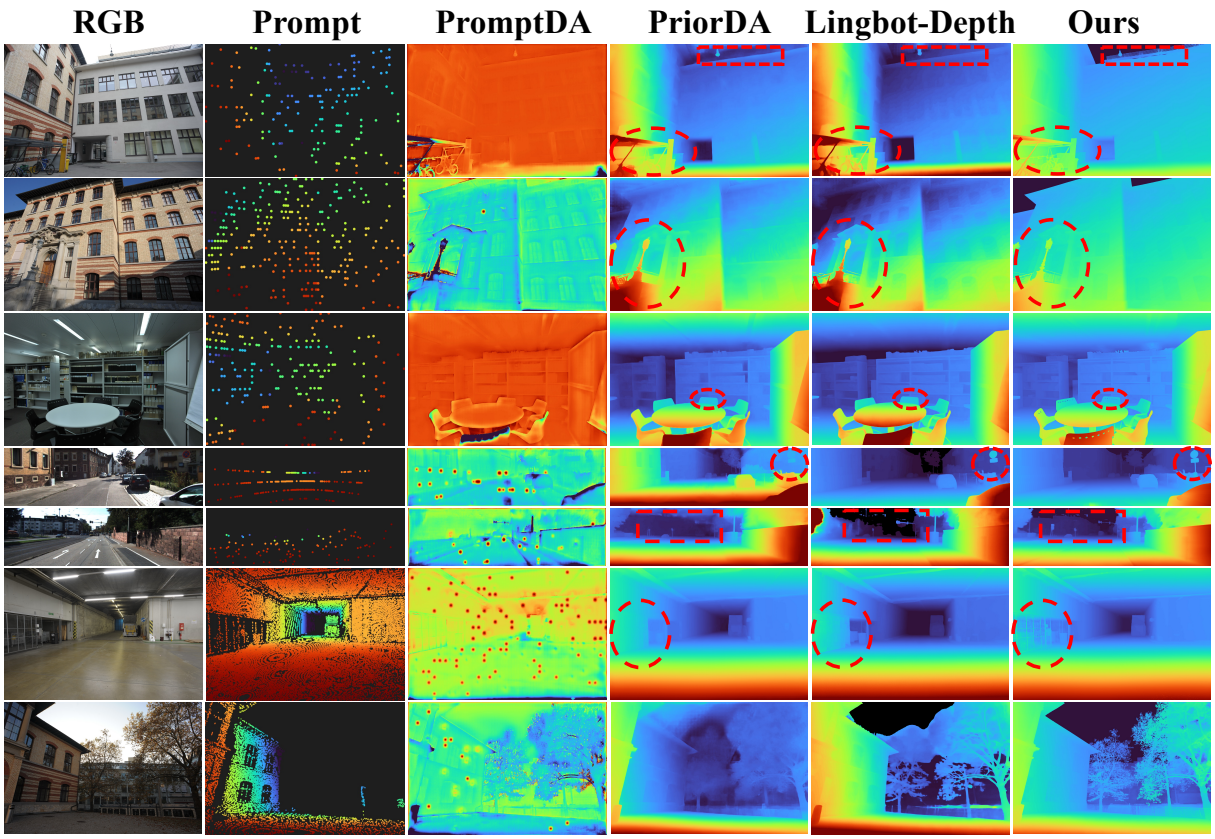


Figure 5: **Visualization of Depth SR and Completion.** Our method better recovers missing regions with improved structure.

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

	NYUv2 ↓				ETH-3D ↓				KITTI ↓			
Method	8×	16×	LiDAR	Extreme	8×	16×	LiDAR	Extreme	8×	16×	LiDAR	Extreme
DAv2 [125]	4.77	5.13	4.85	4.77	6.27	7.38	7.41	6.51	9.54	11.22	8.86	9.25
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
Omni-DC [140]	**1.57**	3.11	2.12	2.63	1.86	4.09	**1.88**	1.98	4.05	8.35	5.27	4.17
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
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
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
PriorDA [114]	1.73	2.79	2.01	**2.01**	2.06	3.91	1.90	**1.61**	4.54	8.20	**4.81**	3.76
LingBot-Depth [99]	2.23	3.05	2.19	**2.03**	**1.18**	**1.54**	13.19	4.12	4.24	7.13	6.55	**3.26**
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**

4.1.2 Radar-Camera Depth Estimation.

Radar sensing has gained increasing attention in autonomous driving and robotics for its low cost, all-weather capability, and long effective range. The task of radar–camera depth estimation aims to recover dense metric depth by fusing RGB frames with mmWave Radar. However, Radar point clouds are extremely sparse and noisy—over a thousand times sparser than LiDAR, and often misaligned spatiotemporally with other onboard sensors, leading to a substantial modality gap from the sparse metric prompts used in our pretraining.

To ensure fair evaluation, our pretraining datasets intentionally exclude all Radar data to avoid information leakage. We therefore use this task to rigorously assess the transferability of our pretrained model to an unseen sensor. We conduct two experiments on the nuScenes dataset [12] using the standard split [113, 67, 54, 62, 66, 96, 93]: (1) training the teacher model from scratch using Radar signals as prompts and LiDAR maps as ground truth; and (2) finetuning the pretrained teacher model under the same supervision. For comparison, we include several

Table 2: Radar-Camera Depth Estimation. Comparisons on the nuScenes [12] dataset (in millimeters)

Type	Method	0-50m		0-70m		0-80m	
		MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Independent	Lin *et al.* [62]	2034.9	4316.5	2294.7	5338.2	2371.0	5623.0
	RC-PDA [67]	2225.0	4156.5	3326.1	6700.6	3713.6	7692.8
	R4Dyn [29]	–	–	–	–	–	6434.0
	DORN [65]	1926.6	4124.8	2380.6	5252.7	2467.7	5554.3
	Singh *et al.* [93]	1727.7	3746.8	2073.2	4590.7	2179.3	4898.7
	CaFNet [96]	1674.2	3674.5	2010.3	4493.1	2109.8	4765.6
	Li *et al.* [55]	1524.5	3567.3	1822.9	4303.6	1927.0	4609.6
	TacoDepth [113]	**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 [113]	**1046.8**	**2487.5**	**1347.1**	**3152.8**	**1492.4**	**3324.8**
Metric Anything-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**

Radar–image fusion baselines, including Lin et al. [62], RC-PDA [67], R4Dyn [29], DORN [65] , Singh et al. [93], CaFNet [96], Li et al. [55], RadarCam-Depth [54] and TacoDepth [113].

As shown in Tab. 2, finetuning our pretrained model with Radar prompt achieves state-of-the-art performance—nearly doubling the accuracy of its from-scratch counterpart and surpassing all prior fusion methods. These results demonstrate that our pretraining paradigm scales effectively: training with randomly sampled sparse prompts allows the model to learn versatile metric representations, enabling seamless adaptation to new and unseen sensing modalities such as mmWave Radar. In contrast to previous works that focus on training and testing with single sensors, our experimental results suggest that future research should pay more attention to general pretraining paradigms.

4.2 Prompt-Free Downstream Tasks

4.2.1 Monocular Depth Estimation.

Metric prediction can be broadly divided into two tracks: (1) Monocular metric depth map estimation and (2) Monocular geometry estimation. The first track aims to predict a dense point map where each pixel encodes the metric distance from the camera origin to the visible surface point along its viewing ray. The second track is monocular geometry estimation, which focuses on predicting a per-pixel 3D point map in the camera coordinate system.

Monocular metric depth map. For metric depth map prediction, we design a dedicated student network (Sec.3.3) and train it from scratch, denoted as 'Student-DepthMap' in Tab. 3 and Tab. 4. For comparison, we evaluate our method against a wide range of recent approaches, including Depth Anything [124], Depth Anything V2 [125], Metric3D [127], Metric3D V2 [38], PatchFusion [58], UniDepth [75], ZoeDepth [8], and DepthPro [10]. We adopt the δ_1 metric, which measures the proportion of inlier pixels based on relative error thresholds, and report results on six datasets — Booster [80], Middlebury [88], Sun-RGBD [94], ETH3D [90], NuScenes [12], and Sintel [11]. These datasets cover indoor/outdoor scenes, cartoon or game scenes, street views and extreme situations such as severe weather and lighting conditions. Notably, none of them are used for training, validation, or hyper-parameter tuning — all results are evaluated under a strict zero-shot protocol. Comprehensive evaluation metrics, including AbsRel [49], Log_{10} error, δ_2/δ_3 thresholds, and RMSE are provided in the Tab. 4. Quantitative results are presented Fig. 6 and Fig. 7.

To ensure a strict zero-shot protocol, we exclude any model–dataset pair if the evaluation set was used during model training. Performance is measured using standard depth estimation metrics: AbsRel, $\log_{10}$, and threshold accuracy scores (δ_1, δ_2, δ_3). As reported in Tab. 3 and Tab. 4, our method exhibits remarkable robustness and consistency across diverse domains:

- **Indoor & Complex Geometry:** On **Sun-RGBD** and **Middlebury**, which feature cluttered indoor scenes and high-resolution details respectively, our method achieves the best performance (Rank 1), significantly outperforming Metric3D v2 and Depth Pro (e.g., AbsRel **0.085** vs. 0.156 on Sun-RGBD). This indicates our model's superior capability in resolving intricate local geometry in constrained spaces.

Table 3: Monocular depth estimation performance of our Student-DepthMap against prior methods, measured by δ_1 accuracy (%, ↑).

Method	ETH3D	Booster	NuScenes	Sun-RGBD	Sintel	Middlebury	Rank ↓
DepthAnything [124]	9.3	52.3	35.4	85.0	6.9	39.3	6.00
DAV2 [125]	36.3	**59.5**	17.7	72.4	5.9	37.2	6.17
Metric3D [127]	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**

Table 4: **Monocular Depth Estimation on Diverse Unseen Datasets.** Zero-shot evaluation across six benchmarks covering indoor, outdoor, driving, and synthetic domains. Our method consistently ranks **1st** or **2nd** across all six datasets.

ETH3D [90]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	1.682	0.380	19.78	31.06
DepthAnything v2 [125]	0.370	0.173	64.66	86.26
Metric3D [127]	0.859	0.240	49.29	57.57
Metric3D v2 [38]	**0.124**	**0.053**	**99.55**	**99.90**
PatchFusion [58]	0.256	0.106	88.38	97.31
UniDepth [75]	0.457	0.186	57.67	81.48
ZoeDepth [8]	0.500	0.176	64.45	81.43
Depth Pro [10]	0.327	0.193	61.31	71.23
Ours-Student-Depthmap	**0.147**	**0.064**	**97.71**	**99.90**

nuScenes [12]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	0.453	0.151	73.88	90.30
DepthAnything v2 [125]	0.614	0.326	31.84	47.27
Metric3D [127]	0.422	0.132	77.22	83.61
Metric3D v2 [38]	0.197	0.080	**93.25**	95.74
PatchFusion [58]	0.392	0.226	48.74	76.04
UniDepth [75]	**0.138**	**0.060**	93.01	**96.42**
ZeroDepth [36]	0.237	0.121	82.60	89.91
ZoeDepth [8]	0.498	0.182	64.95	82.70
Depth Pro [10]	0.287	0.164	73.84	84.25
Ours-Student-Depthmap	**0.152**	**0.063**	**96.56**	**98.26**

Sintel [11]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	3.973	0.559	15.42	27.28
DepthAnything v2 [125]	2.226	0.494	18.70	33.82
Metric3D [127]	1.733	0.387	32.38	44.79
Metric3D v2 [38]	**0.370**	**0.216**	**62.92**	**76.87**
PatchFusion [58]	0.617	0.391	35.52	51.44
UniDepth [75]	0.869	0.301	35.72	57.26
ZeroDepth [36]	0.703	0.491	25.63	37.08
ZoeDepth [8]	0.946	0.392	22.70	44.97
Depth Pro [10]	**0.508**	0.230	**59.25**	**71.14**
Ours-Student-Depthmap	0.792	**0.227**	50.01	70.57

Sun-RGBD [94]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	0.114	0.053	98.81	**99.77**
DepthAnything v2 [125]	0.182	0.070	97.65	99.46
Metric3D [127]	1.712	0.382	27.00	34.12
Metric3D v2 [38]	0.156	0.076	96.35	99.55
PatchFusion [58]	0.466	0.961	60.15	60.65
UniDepth [75]	0.087	0.037	**99.33**	**99.80**
ZoeDepth [8]	0.123	0.053	97.95	99.51
Depth Pro [10]	**0.113**	**0.049**	98.51	99.55
Ours-Student-Depthmap	**0.085**	**0.033**	**99.31**	99.65

Middlebury [88]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	0.273	0.149	69.62	86.06
DepthAnything v2 [125]	0.262	0.141	72.07	90.55
Metric3D [127]	1.251	0.305	37.53	58.73
Metric3D v2 [38]	0.450	0.152	73.32	88.61
PatchFusion [58]	0.250	0.108	87.17	98.15
UniDepth [75]	0.324	0.127	80.05	**99.62**
ZeroDepth [36]	0.377	0.179	67.06	78.95
ZoeDepth [8]	0.214	0.115	77.68	90.86
Depth Pro [10]	0.251	**0.089**	**93.17**	96.40
Ours-Student-Depthmap	**0.200**	**0.082**	**96.17**	**99.93**

Booster [80]

Method	AbsRel↓	Log_{10}↓	δ_2↑	δ_3↑
DepthAnything [124]	0.317	0.114	**79.62**	95.23
DepthAnything v2 [125]	0.315	0.110	76.24	94.28
Metric3D [127]	1.332	0.346	13.07	33.98
Metric3D v2 [38]	0.417	0.140	75.78	92.83
PatchFusion [58]	0.719	0.213	49.39	72.89
UniDepth [75]	0.500	0.166	60.90	89.21
ZoeDepth [8]	0.610	0.195	52.66	75.51
Depth Pro [10]	0.336	0.118	79.43	**96.52**
Ours-Student-Depthmap	**0.282**	**0.100**	**84.11**	**96.83**

- **Outdoor & Driving:** On **nuScenes** and **ETH3D**, our method remains highly competitive, ranking 2nd in absolute error metrics while often achieving the highest accuracy in stricter thresholds (δ_3). However, since these GT datasets inherently contain noise, we consider δ_3 metric to be more reasonable and

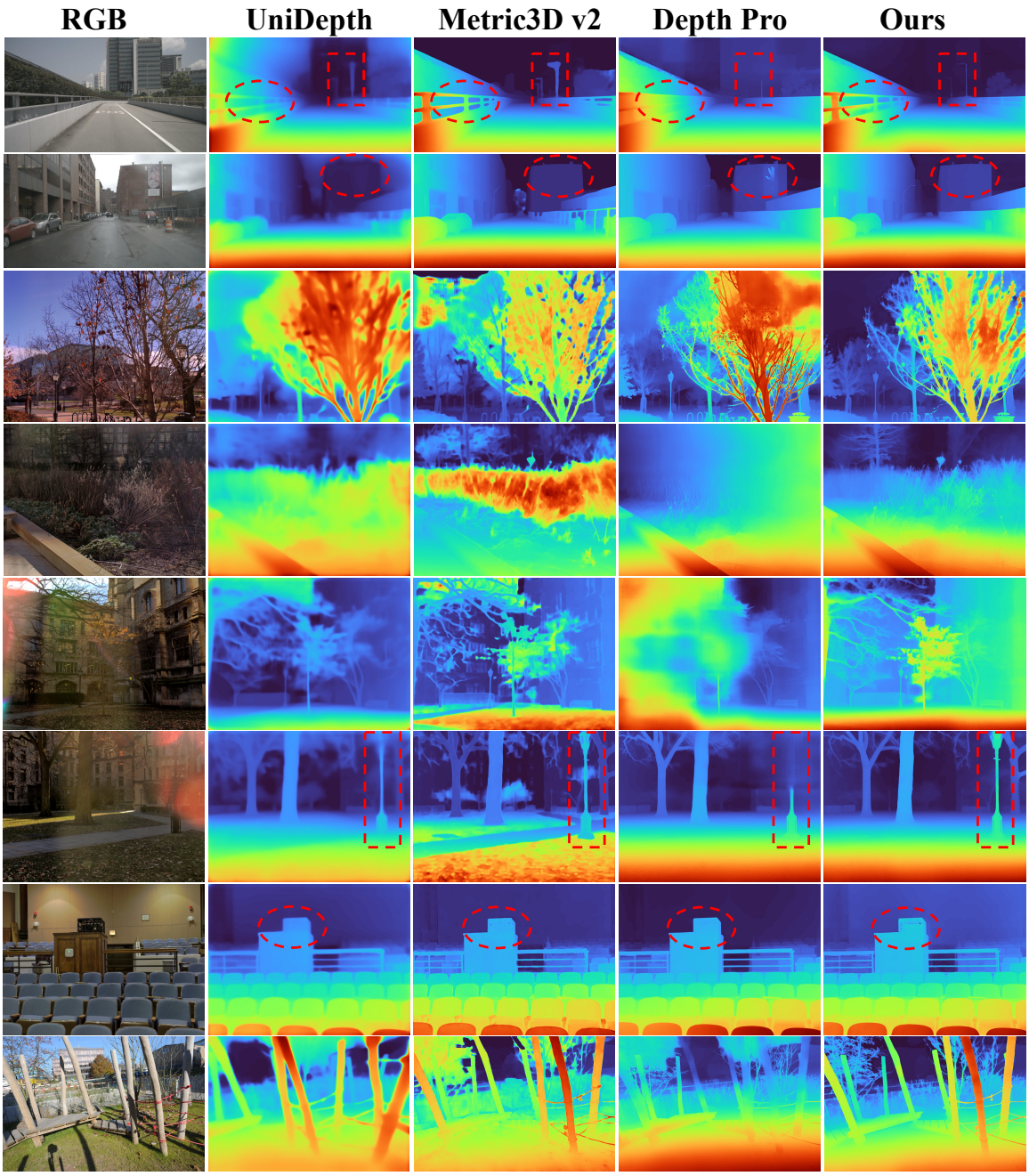


Figure 6: **Zero-shot Visual Comparisons on Challenging Test Samples.** Our model robustly captures details of thin structures and in scenes with difficult lighting where competitors often fail.

robust. This suggests that our model effectively handles large-scale depth variations typical of outdoor environments.

- **Robustness on Unconventional Data:** On **Booster**, a dataset known for challenging lighting and textures, our method outperforms all baselines (AbsRel **0.282**), highlighting its resilience to domain shifts. Even on the synthetic **Sintel** dataset, where domain gaps are significant, we maintain strong performance (2nd best in Log_{10}), demonstrating that our learned representations generalize well beyond photorealistic domains.

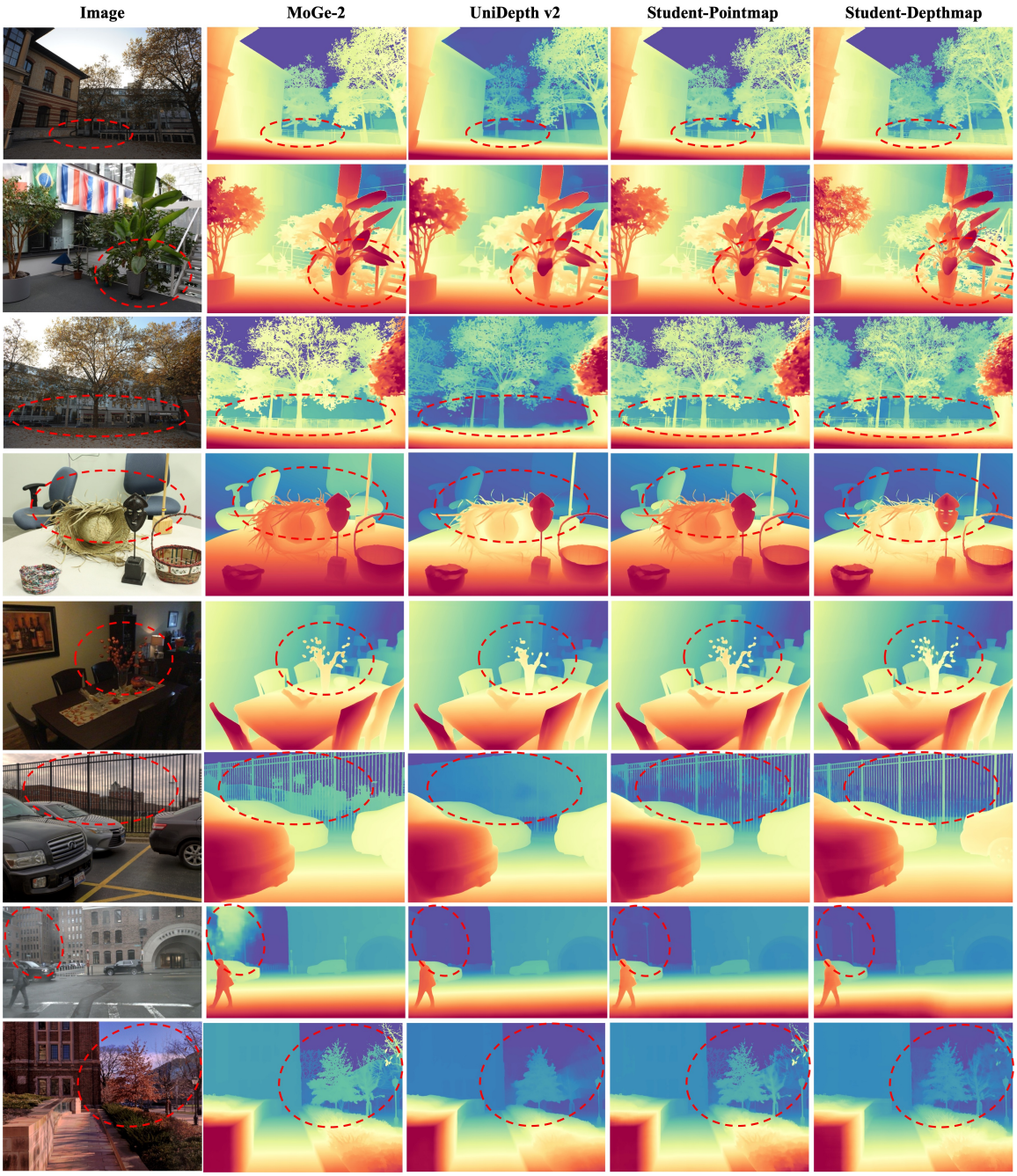


Figure 7: **Qualitative Comparison of Monocular Depth Estimation**. Compared with MoGe2 and UniDepthv2, our distilled model produces more detailed and geometrically plausible predictions for both depth maps and point maps.

Overall, while some baselines excel in specific niches, our method delivers the most balanced and consistently high performance across the full spectrum of test scenarios.

Monocular metric point map. For metric point map prediction, we leverage pseudo-labels generated by our pre-trained model to fine-tune recent state-of-the-art methods such as MoGe-2 [110], denoted as 'Student-PointMap'. This design allows us to assess the generality and precision of our pseudo-labels under different training paradigms — whether training from scratch or fine-tuning, and regardless of whether the output head predicts depth maps or 3D point maps.

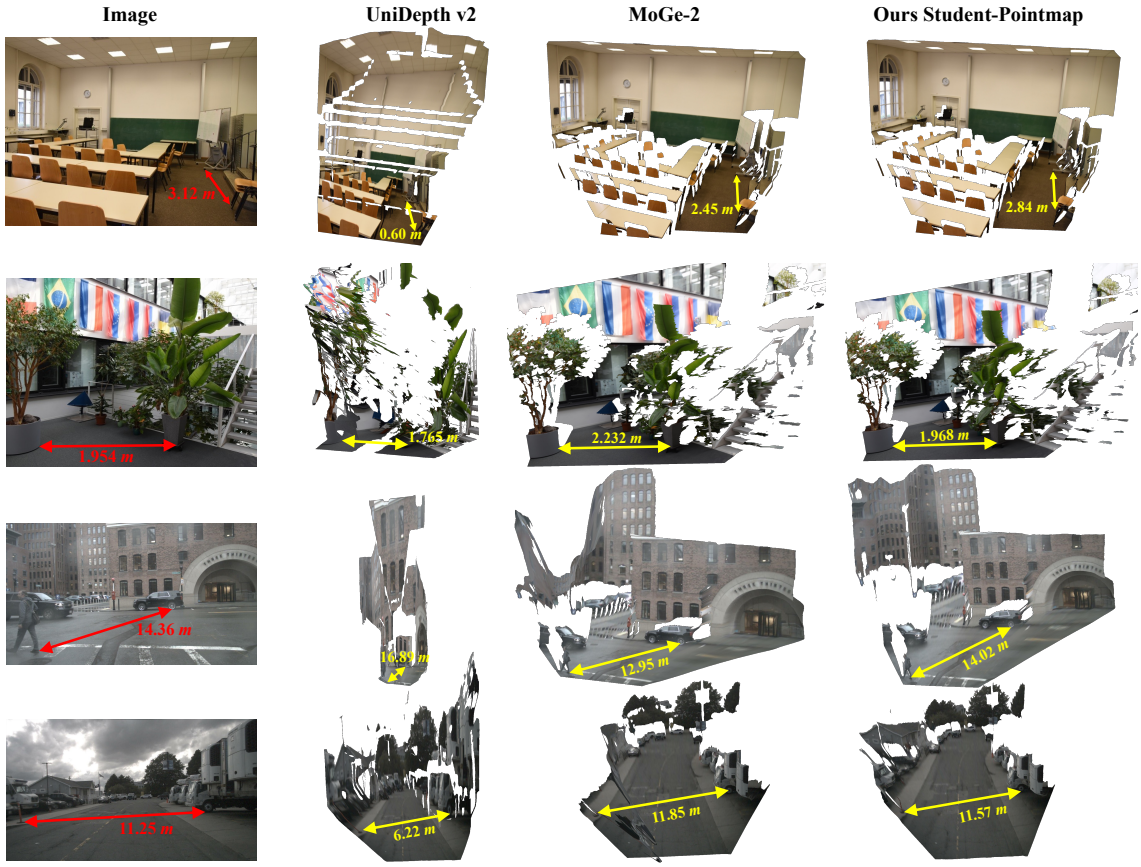


Figure 8: **Qualitative Comparison of Point Maps.** The red arrows indicate the GT distance, the yellow arrows indicate the distance from predicted point map.

Monocular 3D geometry estimation aims to recover a per-pixel 3D point map in the camera coordinate system. In this setting, we leverage pseudo-labels generated by our pre-trained model to fine-tune recent state-of-the-art frameworks such as MoGe-2 [110], denoted as "Student-PointMap". This design enables a comprehensive evaluation of the generality and precision of our pseudo-labels across different training paradigms—including training from scratch versus fine-tuning, and varying output representations (depth maps or 3D point maps). As shown in Fig. 8, Fig. 9, Tab. 5 and Tab. 6, our distillation approach consistently achieves state-of-the-art performance, demonstrating its robustness to differences in prediction heads and network initialization. We adopt the GIANT-LARGE and DA3MONO-LARGE variants from the official DepthAnything3 [59] checkpoints, which represent the largest and most powerful models that support monocular metric depth estimation. However, the performances of GIANT-LARGE and DA3MONO-LARGE are not particularly satisfactory in our setting. We conjecture that the capability of DepthAnything3 [59] still depends on inferring matching relationships across multiple views, making accurate metric scale recovery particularly challenging in complex monocular settings. Additionally, Depth Anything3 [59] does not support inference when the camera's intrinsic parameters are unknown, which limits its applicability. In contrast, we evaluated its performance under both known and unknown intrinsic parameter conditions.

4.2.2 Recovering Camera Intrinsics.

Furthermore, we utilize the point map X predicted by our finetuned model ('Student-PointMap') to infer the intrinsic parameters of the camera from a straightforward optimization. Throughout our experiments, we assume a unit aspect ratio and that the principal point is approximately centered in the image; therefore, the only unknown intrinsic parameter is the focal length of the first camera, denoted f. We estimate f by minimizing a weighted

Table 5: **Monocular Geometry Estimation without Camera Intrinsics.** Our *Student-PointMap* model achieves the best average rank (1.88), demonstrating superior robustness under unknown camera parameters.

Method	KITTI [31]		ETH3D [90]		iBims-1 [48]		DIODE [104]		Avg.		
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Rank↓
ZoeDepth [8]	17.0	85.4	57.1	33.7	17.4	67.2	39.3	29.3	32.70	53.90	7.00
MASt3R [52]	56.7	9.84	47.2	20.1	18.7	61.5	54.9	19.0	44.38	27.61	9.13
DAV1 [124]	11.6	94.5	40.2	24.0	12.9	81.8	58.0	16.2	30.68	54.13	6.88
DAV2 [125]	10.6	88.6	36.1	36.3	**11.1**	**91.7**	41.2	22.1	24.75	59.68	4.88
UniDepth V1 [75]	**4.69**	**98.4**	56.9	14.9	23.8	57.6	**17.1**	**71.9**	25.62	60.70	5.50
UniDepth V2 [74]	8.58	95.4	20.7	69.5	**9.52**	**93.2**	43.0	51.8	20.45	**77.48**	**3.63**
Depth Pro [10]	23.5	38.3	38.5	32.8	15.9	81.5	31.9	37.7	27.45	47.58	6.75
MoGe-2 [110]	18.1	62.9	**10.4**	**90.8**	13.6	83.0	17.5	66.4	**14.90**	75.78	4.13
DAV3-Metric-Large [59]	11.8	83.0	11.0	90.1	27.6	43.5	24.2	53.7	18.66	67.59	5.88
DAV3-Nested-G-L [59]	83.2	0.00	80.7	0.00	39.1	20.9	51.9	26.4	63.73	11.84	10.38
Ours-Student-Pointmap	**3.22**	**97.7**	**10.8**	**90.2**	11.3	86.6	**13.9**	**79.8**	**9.81**	**88.58**	**1.88**

Table 6: **Monocular Geometry Estimation with Provided Camera Intrinsics.** When ground-truth intrinsics are provided, our *Student-PointMap* model significantly outperforms the baseline metric depth estimation models.

Method	KITTI [31]		ETH3D [90]		iBims-1 [48]		DIODE [104]		Avg.		
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Rel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	Rank↓
Metric3D V2 [38]	**5.25**	98.0	11.8	88.8	9.96	**94.1**	49.1	1.98	19.03	70.72	3.38
UniDepth V1 [75]	**4.43**	**98.5**	44.5	26.7	22.6	60.5	21.0	63.5	23.13	62.30	3.75
UniDepth V2 [74]	5.98	97.7	15.0	85.2	**7.71**	**95.5**	41.0	67.1	17.42	86.38	3.13
MoGe-2 [110]	8.64	93.7	**10.5**	**92.2**	9.92	92.4	**16.2**	**77.1**	**11.32**	**88.85**	**2.88**
Ours-Student-Pointmap	5.47	**99.6**	**9.75**	**92.4**	**9.47**	92.0	**14.1**	**74.3**	**9.69**	**89.56**	**1.88**

reprojection error,

$$f^* = \arg\min_f \sum_{i=0}^{W} \sum_{j=0}^{H} \left| (i', j') - f \frac{(X_{i,j,0}, X_{i,j,1})}{X_{i,j,2}} \right|, \tag{6}$$

where i' and j' are the pixel coordinates expressed relative to the image center, and $X_{i,j,k}$ denotes the k-th component of the point map at location (i, j). The optimization in (6) is efficiently solved with a few iterations of a Weiszfeld-type algorithm [115], yielding the estimated focal length f^*. We compare our model to optimizated methods using point map [75, 111, 128] and learning-based camera calibration [43, 138]. As shown in Tab. 7, we achieve the best average accuracy in terms of mean error and median error in degrees.

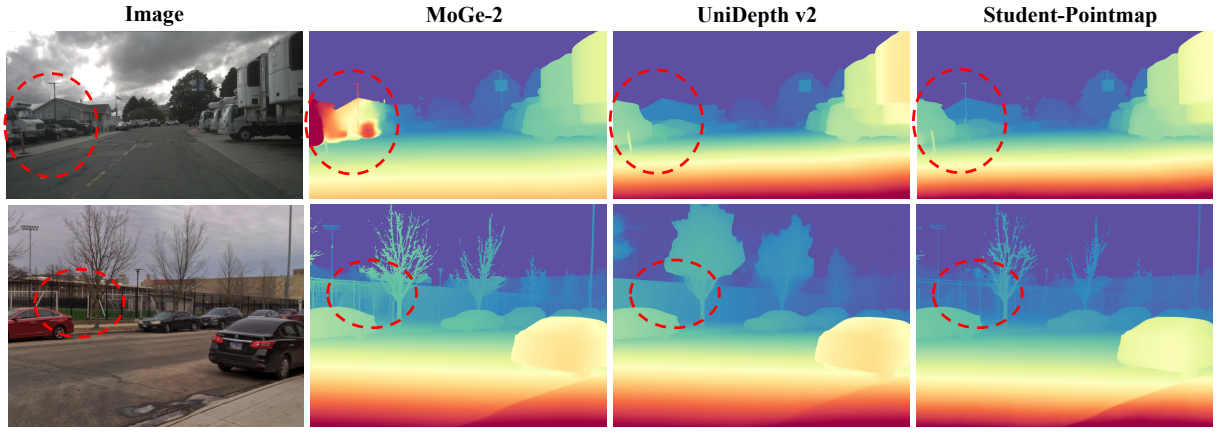


Figure 9: **Qualitative Comparison of Depth Map**. Compared with MoGe2 and UniDepthv2, our Student-PointMap model, which is finetuned from MoGe2 using pseudo-labels predicted by the our pretrained model, achieves more stable and accurate depth estimation.

Table 7: Evaluation Results for Camera Calibration in Degrees.

Method	ETH3D [90]		iBims-1 [48]		Average		
	Mean↓	Med.↓	Mean↓	Med.↓	Mean↓	Med.↓	Rank↓
Perspective [43]	13.6	11.9	10.6	9.30	12.1	10.6	7.25
WildCam [138]	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 [128]	8.26	7.19	18.4	17.5	13.33	12.35	7.00
DUSt3R [111]	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 [110]	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**

4.2.3 Zero-shot Boundaries Accuracy Measure

We evaluate the sharpness of the predicted geometry by our 'Student-DepthMap' using two synthetic benchmarks, Spring [69] and Sintel [11], together with the real-world iBims-1 [48] dataset. As reported in Tab. 8, our method achieves the best average accuracy boundaris compared to DepthPro [10] and MoGe-2 [110].

4.2.4 Multi-view Metric 3D Reconstruction.

To validate the accuracy and cross-view consistency of our monocular depth predictions without prompt, we compare with Map Anything [45], a recent state-of-the-art method for multi-view metric 3D reconstruction method. We use the officially released checkpoint without any additional post-processing or finetuning. Our evaluation takes the multi-view images together with the per-frame monocular depth maps predicted by our prompt-free model as input. No cross-frame correction or post-processing optimization techniques (e.g., Bundle Adjustment) is applied, the 3D metric reconstruction results were obtained through a single feed-forward. We report results on the ETH3D [90] and ScanNet [21] test sets, as show in Tab. 9. Compared with the baseline, our approach achieves superior performance, particularly in metric-scale estimation, as show in Fig. 10. The results show that our proposed pretraining paradigm can transfer its capabilities to a prompt-free student model via distillation. In a prompt-free setting, our model exhibits strong cross-view-consistent depth prediction without relying on any additional conditions. It delivers virtually cost-free improvements in multi-view 3D metric reconstruction accuracy, relying solely on powerful monocular priors rather than multi-view matching-based inference.

Table 8: **Zero-shot Depth Boundary (F1 score [10]) on Multiple Benchmarks.**

Method	Sintel F1↑	Spring F1↑	iBims F1↑	Rank↓
Relative				
DepthAnything [124]	0.261	0.0109	0.127	7.33
DepthAnything v2 [125]	0.228	0.0610	0.111	8.00
Marigold [44]	0.068	-	0.149	7.50
Absolute				
DPT [81]	0.181	-	0.113	9.00
Metric3D [127]	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 [110]	0.282	0.0890	**0.194**	3.67
Student-DepthMap	**0.382**	**0.1635**	**0.179**	**1.67**

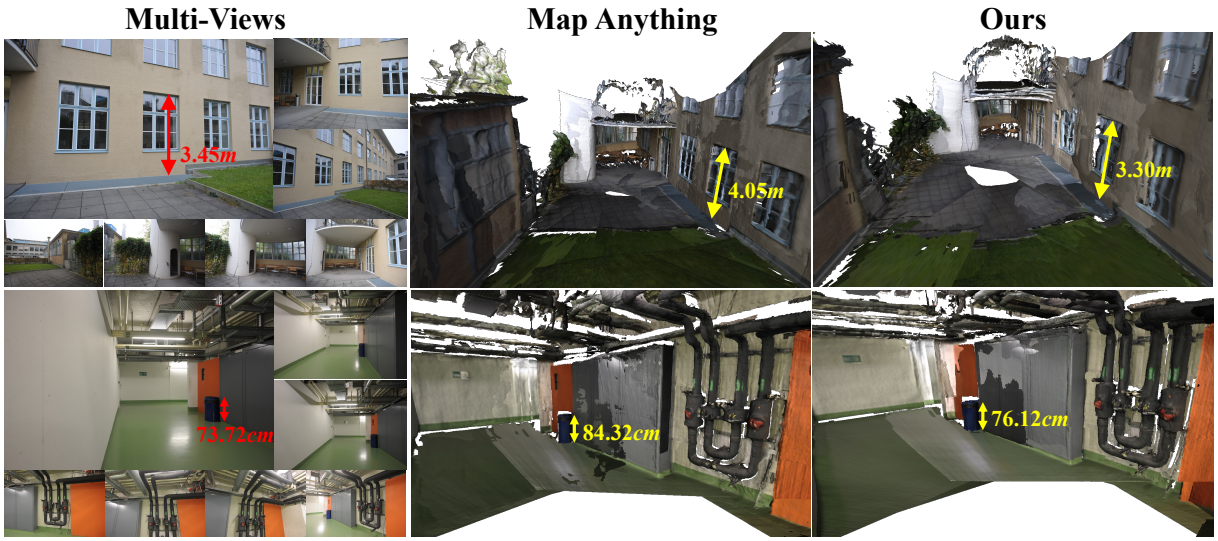


Figure 10: Auxiliary monocular depth inputs improve performance of MapAnything. The red arrows indicate the GT distance, the yellow arrows indicate the distance from 3D reconstruction.

.

Table 9: Integrating our Metric Anything Student module into the MapAnything baseline yields significant performance gains on multi-view metric 3D reconstruction (AbsRel, δ_1 in %).

Method	ETH3D [90]		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**

4.2.5 VLA Planning

We distill our Metric Anything's capability into Vision-Language-Action (VLA) models for action planning. Results (Tab. 10) demonstrate SoTA performance. 3D spatial perception is essential for manipulation in the physical world. Prior work [9] attempt to input depth maps from depth sensors or offline depth-estimation models, along with RGB observations, into vision-language-action (VLA) policies model to improve performance. However, such approaches still require additional hardware or depth predictors at test time, which increases deployment complexity on real robots. Following MolmoACT [51], we distill the depth-perception capability of our trained prompt-free model into the VLA model by supervising it to predict depth tokens, rather than consuming depth

Figure 11: **Enhancing VLA Planning with Metric Anything.** We distill the depth-perception capability of Metric Anything into the VLA model by supervising it to predict metric-aware depth tokens.

Table 10: LIBERO benchmark success rates by task category.

Method	Spatial	Object	Goal	Long	Avg
TraceVLA [137]	84.6	85.2	75.1	54.1	74.8
Octo-Base [102]	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 [135]	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 [73]	**96.4**	**96.8**	88.6	60.2	85.5
ThinkAct [39]	88.3	91.4	87.1	70.9	84.4
Baseline-DAV2 [125]	87.0	95.4	87.6	77.2	86.6
Ours	88.6	94.4	**88.8**	**78.8**	**87.7**

inputs. We adopt Depth Anything V2 [125] as a baseline and report success rates on four tasks from the LIBERO benchmark [63]. As shown in Tab. 10, our model achieves more accurate spatial understanding of both the environment and the target objects than Depth Anything V2, yielding the best average success rate. Moreover, it achieves significant improvements over models without depth perception or input, indicating that distilling depth perception into VLA policies is a promising avenue for enhancing manipulation performance, even without depth inputs during either training or inference. An overview of the framework is provided in Fig. 11

4.2.6 Spatial Understanding of MLLMs

We further evaluate how Metric Anything enhances the 3D spatial reasoning ability of foundation VLMs. Metric Anything encodes rich metric 3D information, and we use its pretrained ViT encoder as a spatial perception backbone to provide 3D-aware features for a vision-language model. Concretely, we extract 3D feature tokens from the ViT encoder of Metric Anything and fuse them with the 2D visual tokens of the VLM via a cross-attention module, where the 2D tokens serve as queries and the 3D tokens as keys and values. In this way, the visual stream of the VLM is explicitly conditioned on the metric 3D prior learned by Metric Anything. The fused tokens are then passed through two linear layers to obtain the final fused representation, which is used as the new visual token input to the language model. Following VLM-3R [9], we adopt LLaVA-Next-Video-7B as the base VLM and conduct supervised fine-tuning on 200K general question-answer pairs, and 4,225 embodied route planning instances released by VLM-3R [9].

Table 11: **Evaluation Results on VSI-Bench [123]**. For Spatial-MLLM and Qwen2.5VL-series [3], we use 16 frames as input. For other open-source methods and GPT-4o [42], we follow the setting of VSI-Bench to set frame numbers (ranging from 8 to 32 frames). For Gemini-1.5 Pro [101], it samples video frames at 1 FPS. **Bold** denote the best-performing open-source models.

Methods	Numerical Question				Multiple-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 [101]	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 [121]	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 [131]	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 [134]	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 [119]	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**

Figure 12: **Enhancing 3D Spatial Reasoning with a Frozen ViT from Metric Anything .** We evaluate our approach on the VIS Benchmark, covering video question-answering tasks like estimating object size, object's distances, appearance order, route planning, and room size. Compared to mainstream large models, our method demonstrates robust and superior performance in 3D spatial understanding.

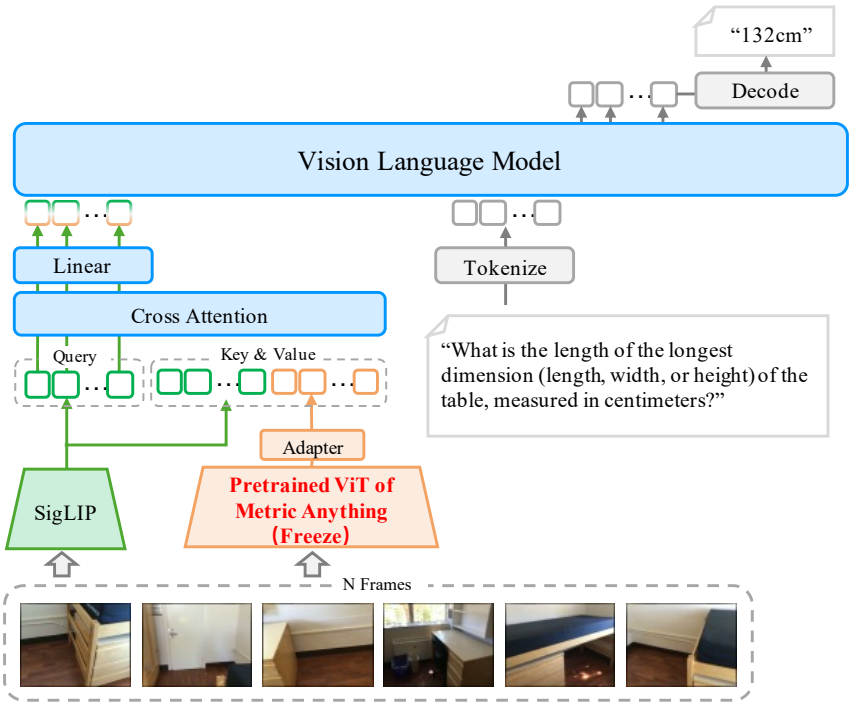


Figure 13: **Enhancing 3D Spatial Reasoning in MLLMs.** We enhance VLM capabilities by employing the frozen, pretrained ViT from Metric Anything as the visual encoder, thereby preserving its strong spatial understanding during fine-tuning.

We report the evaluation results on VSI-Bench [123] in the Tab.11. For Spatial-MLLM and Qwen2.5VL-series [3], we use 16 frames as input. For other open-source methods and GPT-4o [42], we follow the setting of VSI-Bench to set frame numbers (ranging from 8 to 32 frames). For Gemini-1.5 Pro [101], it samples video frames at 1 FPS. As shown in Fig.12, we demonstrate various video question-answer categories within the spatial reasoning benchmark (VIS Benchmark), including object size, inter-object distances, object appearance order, route planning, and room size. Quantitative results and case studies demonstrate that the features extracted by our pretrained model encode rich and accurate spatial understanding, and can substantially enhance the spatial reasoning capabilities of existing VLMs. Since spatial understanding is a core competence for tasks such as embodied intelligence, navigation and path planning, and 3D real-world manipulation, we believe that our proposed pretraining paradigm has the potential to serve as a fundamental visual representation backbone that raises the performance ceiling of these tasks. An overview of the framework is provided in Fig. 13.

5 Ablation Study

5.1 Scaling up Data.

We investigate the impact of training data by ablating different proportions of our training dataset. Metric Anything-Pretrain is trained with 5%, 10%, 20%, 40%, 80%, and 100% of the training set, and its depth super-resolution on 8× downsample performance is evaluated on KITTI, achieving AbsRel values of 5.22, 4.25, 3.98, 3.02, 2.63, 2.34, respectively. The results show that while our performance on small-scale datasets lags behind PromptDA, which uses sophisticated rules to simulate low-resolution prompts, our model achieves the best performance as data scales up due to increasing prompt diversity from varied data sources. By minimizing design biases in prompt, our paradigm offers the most robust generalization to downstream tasks without task-specific pre-design. The data proportion ablation for student is shown in Fig. 2a.

5.2 Network Architecture.

Excluding the benefits of data scaling, we evaluate the effectiveness of our ViT-encoder–DPT-head skip-connection design by comparing it with a classic U-Net–style skip-connection architecture [81]. For both architectures, we

Table 12: Network architecture ablation on NuScenes and ETH3D.

Lable Source	NuScenes [12]				ETH3D [90]			
	Real-world		Pseudo		Real-world		Pseudo	
	Abs↓	RMSE↓	Abs↓	RMSE↓	Abs↓	RMSE↓	Abs↓	RMSE↓
Unet-sytle	0.213	9.631	0.187	7.222	0.327	5.353	0.269	1.44
Ours	0.235	9.662	**0.125**	**6.267**	0.334	5.490	**0.182**	**1.898**

Table 13: Single-frame inference latency under FP32 (ms).

Method	Para.	Native Res.	FLOPs@HD	Latency (ms) @FP32		
				VGA	HD	4K
DepthPro [10]	952M	1536×1536	8848.5G	246.0	246.0	246.0
Ours-PreTrain	993M	1536×1536	19992.6G	285.2	285.2	285.3
Student-PointMap	326M	840×840	4015.7G	29.4	369.8	5043.9
Student-DepthMap	877M	1536×1536	21516.4G	278.8	278.5	278.8

train under two regimes: (1) on noisy, real-world data drawn from multiple sources, and (2) on high-accuracy, domain-aligned pseudo labels produced by our pre-trained model. Results in Tab. 12 show that classic U-Net skip connections perform well on heterogeneous data but fail to fully leverage ViT's semantic representations when trained with consistent pseudo labels. Our Inverse Skip-Connection design better realizes the potential of pseudo-label distillation. Additionally, even with identical architectures, training with our pseudo labels improves performance, demonstrating our pre-trained model's strong spatial understanding.

5.3 Runtime.

To assess the latency of our pretrained model and prompt-free variants in comparison to baselines, we evaluated all approaches at three image resolutions—VGA (640×480), HD (1920×1080) and 4K (4032×3024), and recorded the per-image runtime w/wo prompts. All models were run under FP32 precision in the same environment measured on an H200 GPU. We also report parameter counts and FLOPs for comparison. As summarized in Tab. 13, compared to Depth Pro, which shares the most similar in network architecture to ours, our pretrained model does not exhibit a noticeable increase in inference time despite introducing an additional prompt branch, owing to our design that unifies the image encoder and patch encoder into a shared ViT. We further provide two prompt-free variants, including Student-DepthMap and Student-PointMap. Due to their substantially different parameter counts, their inference latencies vary accordingly. Additional results, including boundary evaluation, hyperparameters, training details and loss ablations are provided in the Appendix.

5.4 Test Time Resolution Scaling

Our metric depth estimation model exhibits a remarkable capability termed "test-time resolution scaling." This allows the model to process input images at resolutions significantly higher than those encountered during training, which results in progressively refined depth predictions. As illustrated in Fig. 14, we present depth maps generated at $1\times$, $3\times$, and $9\times$ the base resolution. The results show that increasing the resolution at test time recovers finer structures and high-frequency details, such as thin object boundaries and texture-rich areas. This demonstrates the model's potential for high-resolution depth sensing without the need for fine-tuning.

5.5 Training Objectives

To evaluate the impact of different training objectives, we trained our model using losses defined on depth, log-depth, inverse-depth, and our proposed distance-balanced inverse-depth loss (Eq. 5 in the main manuscript). Quantitative results in Tab. 14 show that our proposed loss achieves performance comparable to the standard inverse-depth loss at close range, while exhibiting superior performance at longer distances. As the depth range increases, the advantage of our method becomes more pronounced. Considering the depth distribution characteristics shown in Fig. 3 in the main manuscript, and given that our pseudo-labels cover a significantly broader depth range, these results validate the design of our proposed loss function for large-range depth estimation.

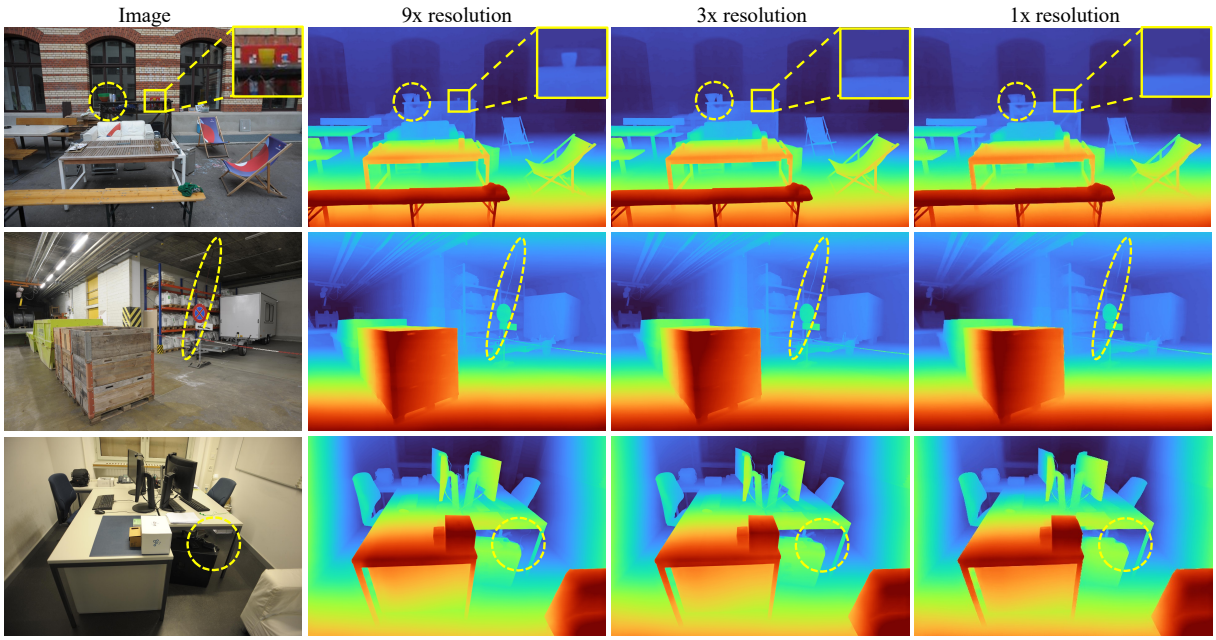


Figure 14: **Test-time Resolution Scaling**. Qualitative results of depth estimation on an example image at $1\times$, $3\times$, and $9\times$ the base input resolution. Higher resolutions recover finer details.

Table 14: **Ablation Study for Different Training Objectives under Varying Depth Ranges**. DIODE [104] AbsRel ↓ metric is reported.

Training objective	DIODE [104] 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**

5.6 Prompt Setting

We analyze the impact of sparse metric-depth prompt density by varying the number of sampled pixels N. Specifically, for each image, we randomly sample 500, 1,000, 2,000, 4,000, 8,000, 16,000, 32,000, or 64,000 valid pixels from the depth map to construct sparse metric prompts. As reported in Tab.15 (left), we present the performance on the Hypersim test set. The results show that as N increases, accuracy gains gradually diminish while computational complexity increases. To balance accuracy and efficiency, we randomly sample $N \in [2{,}000, 40{,}000]$ valid pixels during training and inference. Notably, although we did not explicitly train for extremely sparse prompts (e.g., N = 100), the benefits of data scaling and the diversity of our collected data enable our pretrained model to maintain state-of-the-art performance, as shown in Tab.1 of the main manuscript.

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
AbsRel↓	0.043	0.041	0.038	0.036	0.034	0.033	0.032	0.031
Time (ms)	224	243	256	266	274	284	299	308

C	50	100	200	400	600
0–40 m	0.441	0.452	0.454	0.456	0.465
>40 m	0.635	0.611	0.565	0.562	0.559

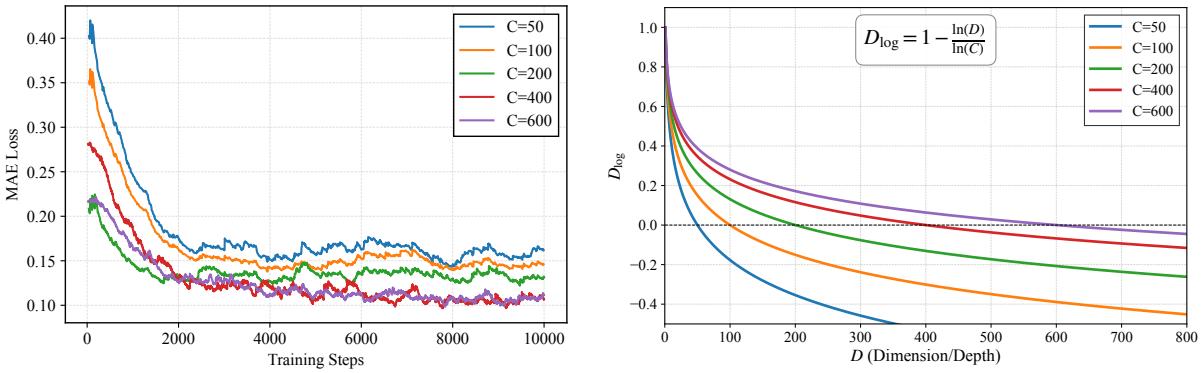


Figure 15: **Analysis of Our Proposed Distance-Balanced Loss.** Left: The training loss for different C values. Right: The loss function curves. (Eq. 5 in the main manuscript).

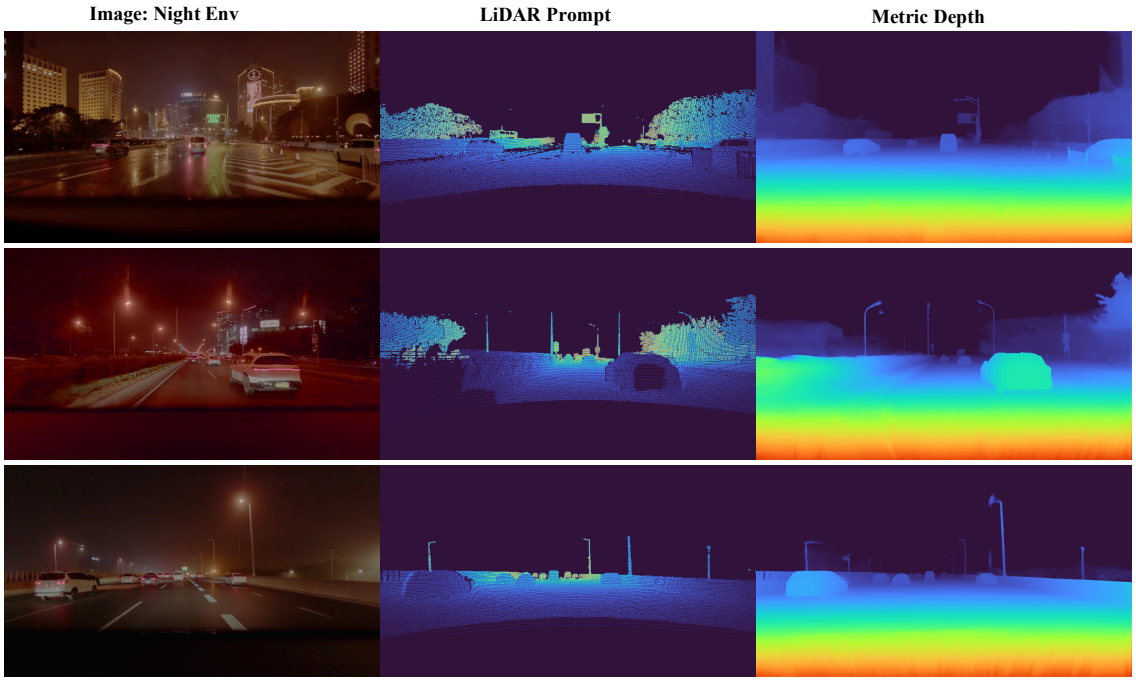


Figure 16: **Robustness in Night-Time Driving**. We deployed a test vehicle to evaluate performance under low-light conditions. As scene brightness drops, visual signals deteriorate and object details fade. Despite this severe degradation, our model maintains remarkably robust.

5.7 Balance Weights

We explore the effect of the balance weight C in Eq.14 in the main manuscript. We try different candidates including $\{50, 100, 200, 400, 600\}$, and report the results in Tab.15 (right). As shown in Fig.15, if the weight is too large, e.g., 600, it tends to encourage the network to learn depth at distant regions while neglecting details in nearby areas. Conversely, when the weight is too small, the network tends to focus on near-field geometric details at the expense of supervision for distant regions. We therefore set this hyperparameter to $C = 400$, which provides a reasonable trade-off.

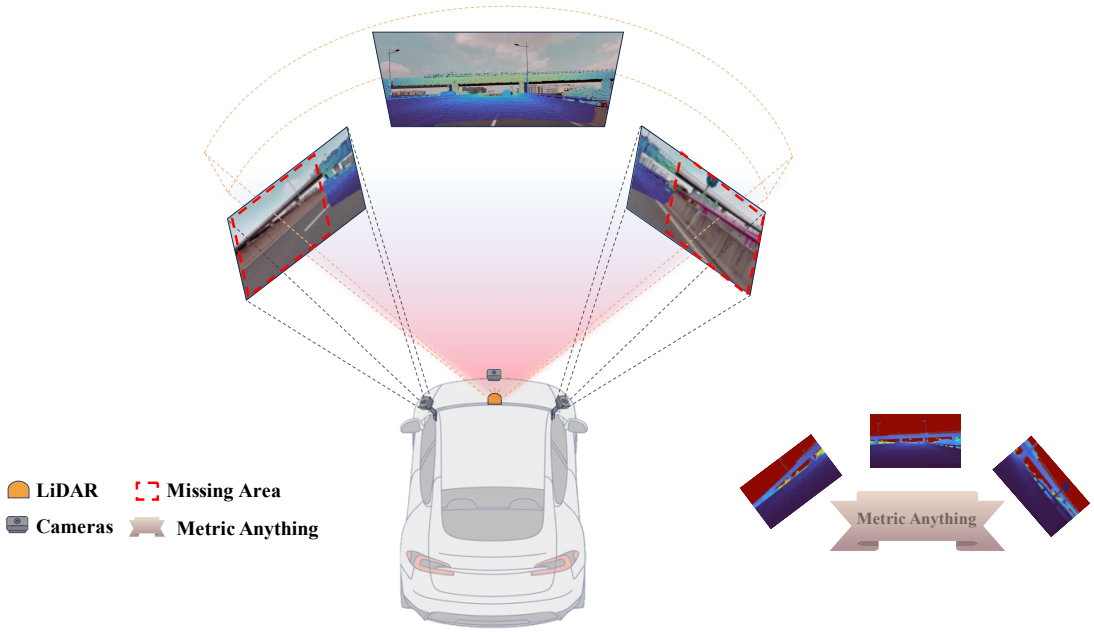


Figure 17: **Sensor Configuration for Real-World Generalization Evaluation.** Our real-world test vehicle is equipped with three cameras (front, left-front, right-front) and a 128-beam solid-state LiDAR. Due to the LiDAR's limited vertical field of view(pitch angle limitation), its captured point cloud does not fully cover the cameras' combined frustums, leaving large image regions without metric depth cues.

6 Generalizability to Unseen Sensors, Scenarios, and Extreme Environmental Conditions

6.1 Generalization across Sensor Configurations

This subsection assesses the model's generalization capability to variations in sensor hardware configuration and data characteristics. We deployed a test vehicle equipped with a sensor suite that differed from the training set in both type and spatial arrangement. The setup consisted of three cameras providing front, right-front, and left-front views, coupled with a 128-beam solid-state LiDAR for forward scene perception (see Fig. 17). The collected real-world data exhibits two key challenges: 1) minor calibration inaccuracies and asynchronous sampling rates—with cameras operating at 24 Hz and LiDAR at 10 Hz—introduced spatiotemporal misalignments between sensor modalities; 2) the LiDAR's field of view did not fully cover the lateral areas captured by the side-facing cameras. We deliberately avoided additional post-processing techniques, such as motion compensation, to rigorously evaluate the model's inherent robustness under these realistic imperfections. The model's performance on two critical tasks is visualized in Fig. 18: **depth completion** for the lateral blind spots (left-front and right-front views) and **super-resolution** for the front view. Together, these results demonstrate that our model can faithfully recover the scene's metric depth even when presented with imperfect, real-world data from an unseen sensor configuration.

6.2 Robustness under Environmental Degradation

This subsection examines the model's robustness under conditions where environmental interference degrades perceptual signals. Two typical scenarios of signal degradation were considered:

- **Night-time driving:** Night-time environments introduce multiple challenges including significantly reduced signal-to-noise ratios, loss of texture and color information, over-saturation from artificial light sources, and high-contrast shadows. These factors substantially impact the reliability of vision-based perception systems.
- **Rainy/Foggy weather conditions:** LiDAR sensors suffer from reflectivity issues that produce anomalous signals or artifacts. This scenario tests whether our model can rely on visual signals to generate reasonable predictions when LiDAR inputs are corrupted.

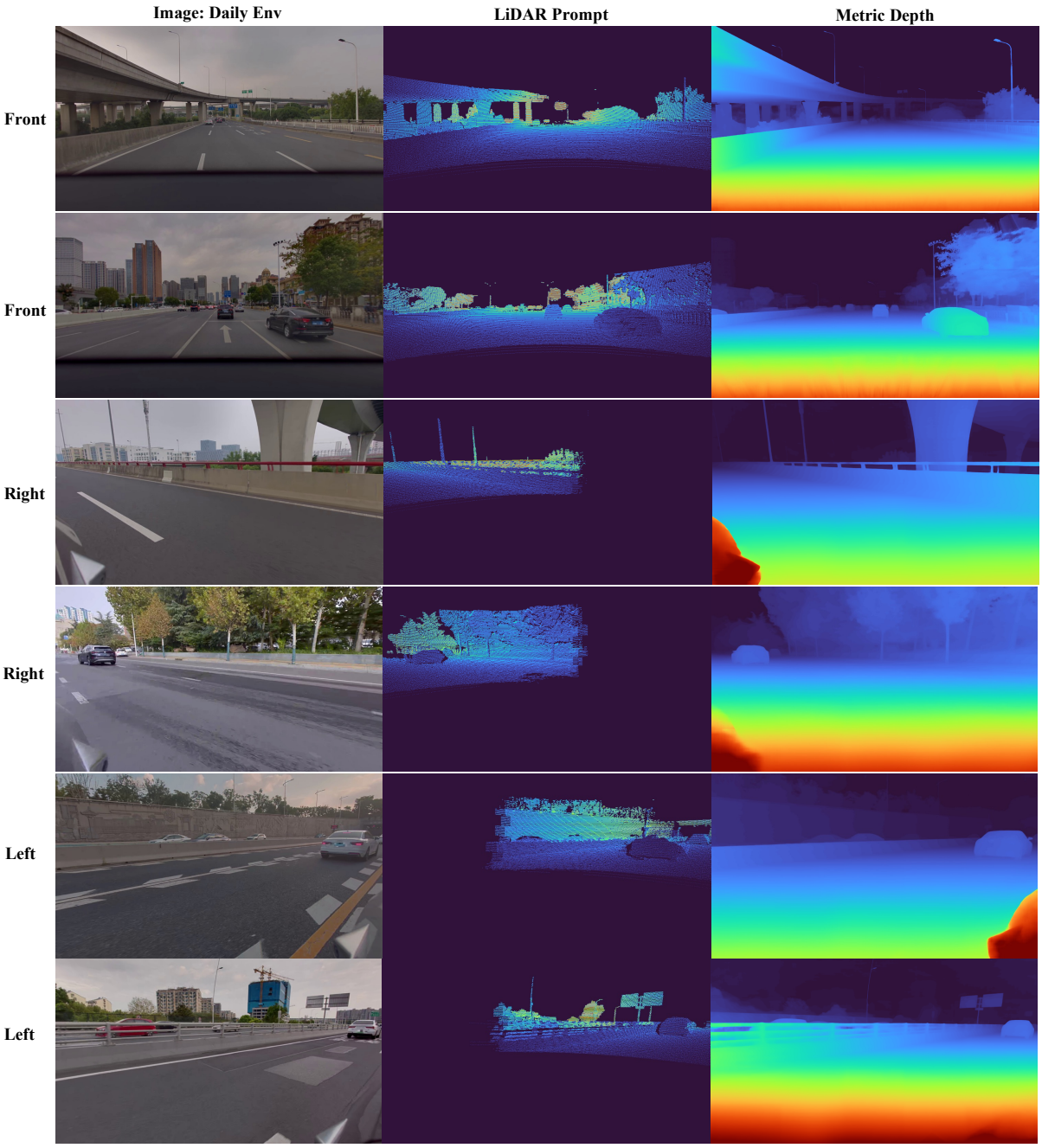


Figure 18: **Generalization to Real-World Sensor Configurations**. We deployed a test vehicle to evaluate in-the-wild depth super-resolution and completion performance of our pre-trained model without any fine-tuning.

As shown in Fig. 19 and Fig. 16, our model maintains reliable depth estimation in both scenarios, demonstrating strong robustness against environmental degradation. **The supplementary video further shows the stability of long-term temporal predictions in our real-world application.**

6.3 Generalization to Unseen Visual Domains

This subsection evaluates the model's zero-shot generalization on monocular depth estimation across visual domains absent from training. Tests were conducted without prompt guidance on three challenging scenarios: **panoramic images** from spherical projections, **fisheye images** with extreme distortions, and diverse **in-the-wild**

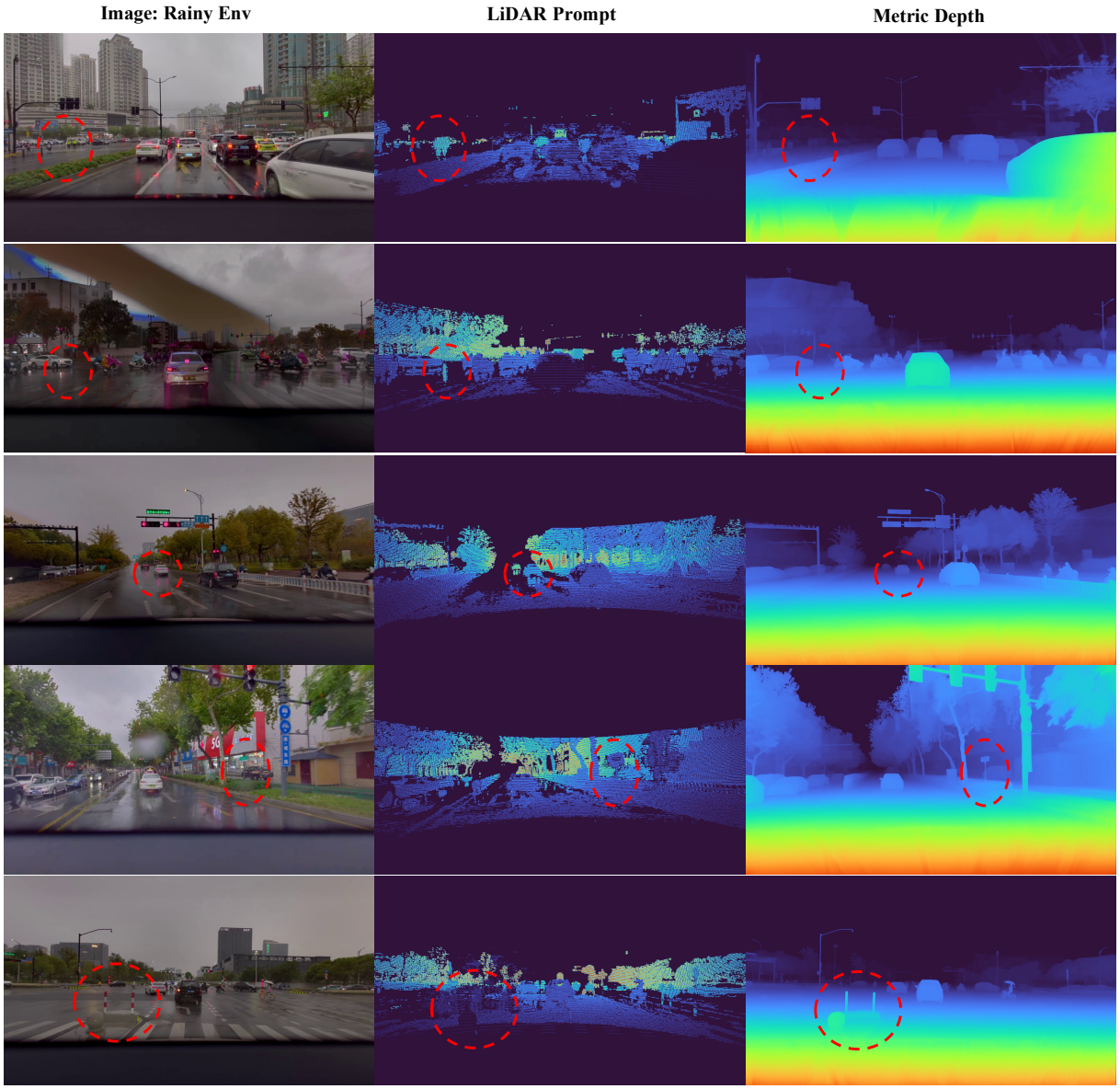


Figure 19: **Robustness in Adverse Weather**. In the real-world deployment, we used a test vehicle to evaluate our pre-trained model for depth super-resolution and completion in rainy and foggy weather conditions without fine-tuning. These adverse conditions significantly affect scene reflectance, causing the LiDAR to produce numerous artifacts or completely occlude critical objects. For example, the degraded data can lead to flat ground surfaces being misinterpreted as uneven or crucial obstacles like pillars being missed. However, our model robustly ignores these erroneous inputs and generates accurate depth predictions based on visual cues, thereby demonstrating the complementary strengths of the two sensing modalities.

scenes including cartoons, grayscale images, and artistic renderings. Qualitative results (Fig. 20, Fig. 21, Fig. 22, and Fig. 23) confirm accurate metric depth estimation throughout. This robust performance across domains previously unrepresented in training data substantiates our claim of achieving "Metric Anything" generalization.

Figure 20: **Generalization to Unseen Visual Domains.** Depth prediction results on **fisheye images**, an unseen domain characterized by severe radial distortion. The model was applied in a zero-shot setting without fine-tuning.

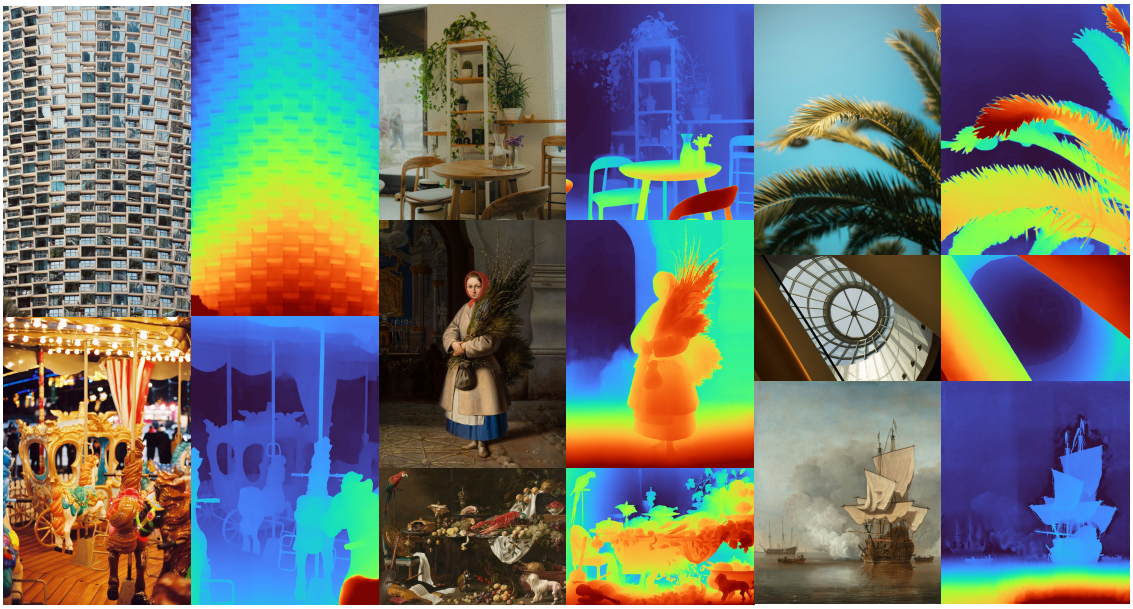


Figure 21: **Generalization to Unseen Visual Domains.** Depth prediction visualization for diverse in-the-wild images.

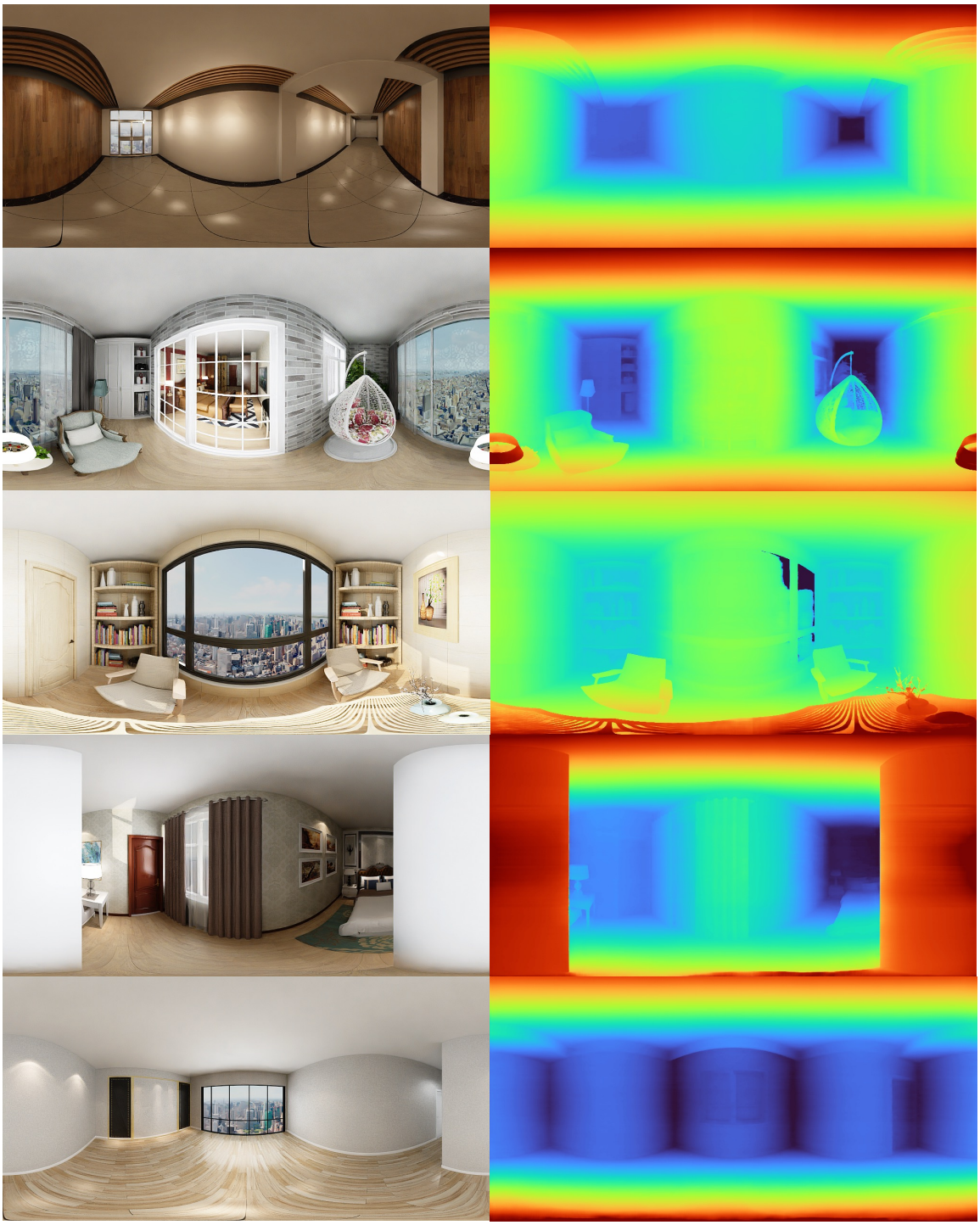


Figure 22: **Generalization to Unseen Visual Domains.** Visualizing depth predictions on **panoramic images**, an unseen domain during training. Our model successfully handles such extreme distortion and novel viewpoints.

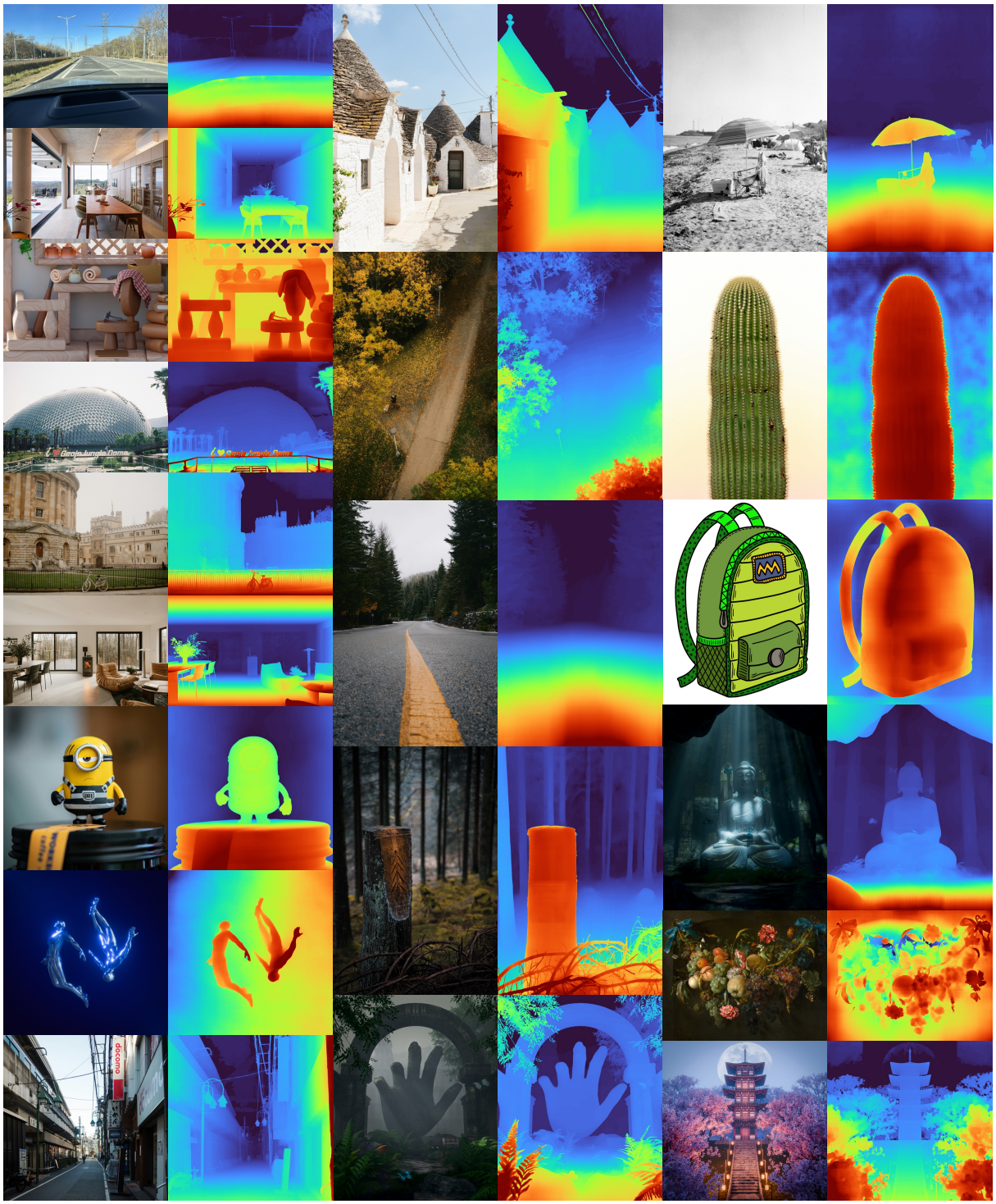


Figure 23: **Generalization to Unseen Visual Domains.** Additional visualizations of depth predictions on diverse in-the-wild images.

7 Training Details

7.1 Training and Test Set Split.

Across all experiments, results are reported in a zero-shot manner, meaning that the training and test sets are sourced from completely distinct datasets with no shared origin. The only exception is ScanNet [21], where we utilize its training set for model training and evaluate on its test set in the multi-view 3D metric reconstruction experiments, following the exact protocol of Map Anything [45]. This usage of ScanNet for training is fair to other methods, as it is also used for training in Map Anything.

7.2 Training Details of Pre-trained Model.

We train our model on 144 H200 GPUs for 100k steps, using a 10k-step warm-up, a peak learning rate of 1×10^{-6} for the ViT backbone, and 1×10^{-5} for the DPT head and the prompt layer. To incorporate the prompt, we introduce the required convolutional kernels and biases with zero initialization.

7.3 Training Details of Distilled Model.

We distill the pre-trained model into a prompt-free student, supervised by pseudo-labels generated through pre-trained model inference. To validate the generality of our pre-trained model, we design two training paradigms: (1) From-scratch training, where we initialize the ViT backbone from DINOv3 ViT-H+/16 [92] and randomly initialize the DPT head. We train our model on 144 H200 GPUs for 200k steps, using a 5k-step warm-up with a peak learning rate of 2×10^{-6} for the ViT backbone and 2×10^{-5} for the DPT head. (2) Fine-tuning state-of-the-art methods: We initialize all network parameters from MoGo-2 [110]. We train our model on 80 H200 GPUs for 10k steps, employing a 1k-step warm-up with a peak learning rate of 2.5×10^{-6} for the ViT backbone and 2.5×10^{-5} for the DPT head.

For all experiments mentioned above, including pre-training and distillation, we supervise the network using the original resolution of depth maps from the datasets. Specifically, inputs are uniformly resized to 1536×1536, and the network outputs are produced at the same resolution. However, for loss computation during training supervision, we resize the network outputs to match the original dataset resolution for comparison with ground truth. To enhance training efficiency, we utilize FlashAttention [22] alongside DeepSpeed with ZeRO Stage-2 optimizer [83] and BF16 precision.

7.4 Training Details of Vision-Language-Model.

Our VLM is initialized from LLaVa-NeXT-Video-7B [133] and equipped with two visual encoders, a SigLIP based image encoder and a our pretrained ViT of Metric Anything. The features of Metric Anything are first mapped to a 1152 dimensional space using a lightweight adapter. We then fuse the SigLIP features and the adapted Metric Anything features with a cross attention module, and apply a projection layer to match the multimodal embedding dimension of the backbone before feeding the fused representation into the language model. The projection layer is initialized from the pretrained LLaVA-NeXT model, while the adapter and cross attention module are initialized from scratch. For training, we adopt the dataset configuration introduced in VLM-3R [9] and fine-tune our VLM for 2K steps with a per-device batch size of 4, using 10 nodes with 8 GPUs per node. During training, the LLM backbone, the vision adapter, the cross-attention fusion modules, and the projection layer are all kept trainable, while all remaining vision encoders are frozen. We use AdamW with a learning rate of 1e-5 for the trainable modules, a cosine learning rate schedule with a 0.03 warmup ratio, and gradient clipping with a threshold of 0.5.

7.5 Training Details of Vision-Language-Action Model.

For each frame, we first extract a depth map using our distilled depth estimation model. The depth maps are then normalized using the 5th and 95th percentiles and subsequently converted into depth tokens using a VQ-VAE, which are used as an extra ground-truth signal to supervise the VLA model in predicting depth tokens, while simultaneously predicting the action token.. We apply parameter-efficient fine-tuning via LoRA, attaching low-rank adaptation modules to all Transformer layers in the model. The model is trained on 10 nodes with 8 GPUs per node, using a global batch size of 160. We use AdamW with a learning rate of 5e-4 and a cosine learning rate schedule.

Evaluation Metrics

Depth Metrics. For the quantitative evaluation of depth estimation, we follow the standard metrics from prior works. Let d_i and $\hat{d}_i$ denote the ground truth depth and the predicted depth for pixel i, respectively. The evaluation is performed over all N valid pixels. Our reported metrics are defined as follows:

- **Threshold Accuracy (δ_i):** The percentage of pixels where the ratio of predicted and ground truth depth falls within a certain threshold:

$$\% \text{ of pixels s.t. } \max\left(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}\right) < \delta,$$

  where $\delta \in \{1.25, 1.25^2, 1.25^3\}$.
- **Absolute Relative Error (AbsRel):**

$$\frac{1}{N}\sum_i \frac{|\hat{d}_i - d_i|}{d_i}$$

- **Root Mean Squared Error (RMSE):**

$$\sqrt{\frac{1}{N}\sum_i (\hat{d}_i - d_i)^2}$$

- **Mean Absolute Error (MAE):**

$$\frac{1}{N}\sum_i |\hat{d}_i - d_i|$$

- **Log10 Error:**

$$\frac{1}{N}\sum_i |\log_{10}(\hat{d}_i) - \log_{10}(d_i)|$$

These metrics are used throughout our experiments. Specifically:

- δ_1: Reported in Sec 4 (Tab. 3, Tab. 9) for zero-shot monocular depth estimation evaluation.
- AbsRel: Reported in Sec 4 (Tab. 1, Tab. 9) for metric depth accuracy evaluation.
- RMSE, MAE: Reported in Sec 4 (Tab. 2) for radar-camera depth estimation evaluation.
- Log_{10}, δ_2, δ_3: Used in comprehensive evaluation tables in the appendix for additional depth accuracy assessment.

Boundary Metrics. Following DepthPro [10], we evaluate boundary sharpness using depth-based and mask-based metrics. For depth maps, we define occluding contours based on pairwise depth ratios between neighboring pixels. Let i, j be the locations of two neighboring pixels. We define an occluding contour c_d derived from a depth map d as:

$$c_d(i,j) = \left[\frac{d(j)}{d(i)} > \left(1 + \frac{t}{100}\right)\right],$$

where $[\cdot]$ is the Iverson bracket, indicating the presence of an occluding contour if the depth differs by more than $t\%$. For all pairs of neighboring pixels, we compute precision (P) and recall (R) as:

$$\text{P}(t) = \frac{\sum_{i,j\in N(i)} c_d(i,j) \wedge c_{\hat{d}}(i,j)}{\sum_{i,j\in N(i)} c_d(i,j)},$$

$$\text{R}(t) = \frac{\sum_{i,j\in N(i)} c_d(i,j) \wedge c_{\hat{d}}(i,j)}{\sum_{i,j\in N(i)} c_{\hat{d}}(i,j)},$$

where $N(i)$ denotes the set of neighboring pixels of i, c_d and $c_{\hat{d}}$ are occluding contours from predicted and ground-truth depth maps, respectively. The **Boundary F1 score** is computed as:

$$\text{F1}(t) = \frac{2 \cdot \text{P}(t) \cdot \text{R}(t)}{\text{P}(t) + \text{R}(t)}. \tag{7}$$

Camera Intrinsics Metrics. For focal length estimation evaluation, we report the mean and median angular errors in degrees. Given the ground-truth focal length f_{gt} (in pixels) and image width w, the horizontal field of view (FOV) is computed as:

$$\text{FOV} = 2 \arctan\left(\frac{w}{2f_{\text{gt}}}\right).$$

The angular error in degrees is:

$$\text{Error} = |\text{FOV}_{\text{pred}} - \text{FOV}_{\text{gt}}| \times \frac{180}{\pi},$$

where FOV_{pred} and FOV_{gt} are predicted and ground-truth FOV values, respectively. FOV estimation results are reported in Sec 4 (Tab. 7) for camera calibration evaluation.

7.6 Loss Functions

Pre-train Model Losses. The pre-train model is trained to predict dense metric depth from monocular images conditioned on sparse metric prompts. We operate in the inverse depth space C. For real-world metric datasets, we adopt a **Robust MAE** loss that discards the top 20% of pixels with the largest errors per image to mitigate the influence of noisy ground truth:

$$\mathcal{L}_{MAE}(\hat{C}, C) = \frac{1}{N'} \sum_{i \in \mathcal{S}} |\hat{C}_i - C_i|, \tag{8}$$

where $\mathcal{S}$ is the set of pixels after removing the top 20% largest errors, and $N' = |\mathcal{S}|$. For synthetic datasets without metric scale, we additionally apply the **Scale-and-Shift-Invariant Mean Absolute Gradient Error (SSI-MAGE)** loss [10]. First, we normalize predictions and ground truth via mean absolute deviation from the median to achieve scale and shift invariance:

$$\tilde{C} = \frac{C - \text{median}(C)}{\text{MAD}(C)}, \tag{9}$$

$$\tilde{\hat{C}} = \frac{\hat{C} - \text{median}(\hat{C})}{\text{MAD}(\hat{C})}, \tag{10}$$

where $\text{MAD}(C) = \text{median}(|C - \text{median}(C)|)$ is the median absolute deviation. Then, we compute the multi-scale gradient loss. Let ∇_S denote the Scharr gradient operator [87]. The multi-scale derivative loss over M scales is defined as:

$$\mathcal{L}_{*,p,M}(C, \hat{C}) = \frac{1}{M} \sum_{j}^{M} \frac{1}{N_j} \sum_{i}^{N_j} |\nabla_* C_i^j - \nabla_* \hat{C}_i^j|^p, \tag{11}$$

where C^j and $\hat{C}^j$ are the inverse depth maps at scale j (obtained by blurring and downsampling by a factor of 2 per scale), and N_j represents the number of valid pixels at scale j. The **SSI-MAGE** loss is defined as:

$$\mathcal{L}_{SSI\text{-}MAGE}(\hat{C}, C) = \mathcal{L}_{S,1,6}(\tilde{\hat{C}}, \tilde{C}), \tag{12}$$

where $\mathcal{L}_{S,1,6}$ denotes the multi-scale gradient loss with Scharr operator (S), L_1 norm ($p = 1$), and 6 scales ($M = 6$). The overall training objective for the teacher model is:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{MAE} + \beta \mathcal{L}_{SSI\text{-}MAGE}, \tag{13}$$

where α and β are weighting coefficients, we set $\alpha = 15$ and $\beta = 5$. For synthetic datasets, both terms are applied; for real-world datasets, only the robust MAE loss is used. This loss formulation is described in Sec 3.2 of the main manuscript.

Distill Model Losses. The distill model is trained on pseudo labels generated by the teacher model. To handle the wide depth range (from near to far distances) in the pseudo labels (as shown in Fig. 3), we design a **Distance-Balanced Inverse-Depth Loss**. The loss operates in a log-space representation that preserves fine-grained sensitivity in near regions while extending effective supervision to long-distance areas. The transformed depth value is defined as (Eq. 5 in the main manuscript): $D_{\log} = 1 - \ln(D)/\ln(C)$, where D is the metric depth and $C = 400$ is a hyperparameter controlling the trade-off between long-range and short-range supervision. The distance-balanced loss is then computed as:

$$\mathcal{L}_{Student}(\hat{D}, D) = \frac{1}{N} \sum_{i}^{N} |D_{\log}(\hat{D}_i) - D_{\log}(D_i)|, \tag{14}$$

where $\hat{D}$ and D are predicted and ground-truth metric depth maps, respectively. This loss function addresses the limitation of standard inverse-depth loss, which decays too rapidly with distance, by providing more balanced supervision across the full depth range covered by our teacher-generated pseudo labels. Additionally, we apply the **SSI-MAGE** loss (Eq. 12) to the log-space transformed depth maps, combining it with our distance-balanced supervision to further enhance boundary sharpness and geometric detail preservation. The overall student training objective is:

$$\mathcal{L}_{StudentTotal} = \gamma \mathcal{L}_{Student} + \delta \mathcal{L}_{SSI\text{-}MAGE}(D_{\log}(\hat{D}), D_{\log}(D)), \tag{15}$$

where γ and δ are weighting coefficients, we set $\gamma = 10$ and $\delta = 2$. The design rationale and experimental validation are presented in Sec 3.3 of the main paper and Tab. 14.

8 Limitations

Our work maintains the central projection camera assumption and has not been extended to specialized camera models (e.g., non-central or non-pinhole configurations). In terms of model scalability, while our data-centric scaling strategy demonstrates strong empirical gains, the scalability of the model architecture itself remains unexplored. Expanding the model's architecture could potentially enhance its capability for depth perception in more complex and diverse scenarios.

9 Conclusion

We present **Metric Anything**, a scalable pretraining framework for metric depth estimation that learns from diverse, noisy 3D sources without task-specific architectures or manually engineered prompts. Using Sparse Metric Prompts to separate spatial reasoning from sensor and camera biases, our approach effectively leverages heterogeneous data. Experiments reveal, for the first time, a clear scaling effect in the metric depth trick. Both the pretrained model and its distilled prompt-free student achieve state-of-the-art results across a wide range of downstream tasks. These results indicate more efficient general-purpose solutions for real-world depth perception.

References

[1] Manuel López Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulo, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *European Conference on Computer Vision*, pages 589–604. Springer, 2020.

[2] Baidu Apollo. Apollo synthetic dataset, 2019. Accessed: 2025-6-12.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.

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

[5] Zuria Bauer, Francisco Gomez-Donoso, Edmanuel Cruz, Sergio Orts-Escolano, and Miguel Cazorla. Uasol, a large-scale high-resolution outdoor stereo dataset. *Scientific data*, 6(1):162, 2019.

[6] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021.

[7] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022.

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

[9] Vineet Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 3d cavla: Leveraging depth and 3d context to generalize vision language action models for unseen tasks. *arXiv preprint arXiv:2505.05800*, 2025.

[10] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025.

[11] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, pages 611–625. Springer, 2012.

[12] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[13] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

[14] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[15] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016.

[16] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. 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] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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*, pages 24185–24198, 2024.

[18] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *AAAI*, 2020.

[19] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *TPAMI*, 2019.

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[21] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[22] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[24] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015.

[25] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

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

[27] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[28] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[29] Stefano Gasperini, Patrick Koch, Vinzenz Dallabetta, Nassir Navab, Benjamin Busam, and Federico Tombari. R4dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 751–760. IEEE, 2021.

[30] Mathias Gehrig, Willem Aarents, Daniel Gehrig, and Davide Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021.

[31] A Geiger, P Lenz, and R Urtasun. Are we ready for autonomous driving. *The KITTI vision benchmark suite. InCVPR*, 2:5, 2012.

[32] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019.

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

[34] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020.

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

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

[37] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *arXiv preprint arXiv:2006.14480*, 2020.

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

[39] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025.

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

[41] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025.

[42] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[43] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F Fouhey. Perspective fields for single image camera calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17307–17316, 2023.

[44] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024.

[45] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction, 2025. arXiv preprint arXiv:2509.13414.

[46] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[47] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

[49] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–96, 2014.

[50] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[51] Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, et al. Molmoact: Action reasoning models that can reason in space. *arXiv preprint arXiv:2508.07917*, 2025.

[52] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[53] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[54] Han Li, Yukai Ma, Yaqing Gu, Kewei Hu, Yong Liu, and Xingxing Zuo. Radarcam-depth: Radar-camera fusion for depth estimation with learned metric scale. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10665–10672. IEEE, 2024.

[55] Huadong Li, Minhao Jing, Wang Jin, Shichao Dong, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Sparse beats dense: Rethinking supervision in radar-camera depth completion. In *European Conference on Computer Vision*, pages 127–143. Springer, 2024.

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

[57] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018.

[58] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. 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*, pages 10016–10025, 2024.

[59] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[60] 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.

[61] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26679–26689, 2023.

[62] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. in 2020 ieee. In *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240, 2020.

[63] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

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

[65] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347. IEEE, 2021.

[66] Chen-Chou Lo and Patrick Vandewalle. Rcdpt: Radar-camera fusion dense prediction transformer. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[67] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan. Radar-camera pixel depth association for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12507–12516, 2021.

[68] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. 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*, pages 2016–2029, 2025.

[69] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. 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*, pages 4981–4991, 2023.

[70] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[71] Jin-Hwi Park, Chanhwi Jeong, Junoh Lee, and Hae-Gon Jeon. Depth prompting for sensor-agnostic depth estimation. In *CVPR*, 2024.

[72] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[73] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[74] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025.

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

[76] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[78] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[79] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[80] Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. 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] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[82] 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, 2020.

[83] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

[84] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

[85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[87] Hanno Scharr, Stefan Körkel, and Bernd Jähne. Numerische Isotropieoptimierung von FIR-Filtern mittels Querglättung. In *DAGM-Symposium*, 1997.

[88] 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 *German conference on pattern recognition*, pages 31–42. Springer, 2014.

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

[90] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[91] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012.

[92] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[93] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. Depth estimation from camera image and mmwave radar point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9275–9285, 2023.

[94] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[95] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[96] Huawei Sun, Hao Feng, Julius Ott, Lorenzo Servadei, and Robert Wille. Cafnet: A confidence-driven framework for radar camera depth estimation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2734–2740. IEEE, 2024.

[97] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[98] Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38, 2019.

[99] Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

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

[101] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[102] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[103] Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8942–8952, 2021.

[104] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[105] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[106] 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, 2024.

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

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

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

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

[111] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

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

[113] Yiran Wang, Jiaqi Li, Chaoyi Hong, Ruibo Li, Liusheng Sun, Xiao Song, Zhe Wang, Zhiguo Cao, and Guosheng Lin. Tacodepth: Towards efficient radar-camera depth estimation with one-stage fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10523–10533, 2025.

[114] Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior, 2025.

[115] Endre Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

[116] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. *CVPR*, 2025.

[117] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023.

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

[119] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *ArXiv*, abs/2505.23747, 2025.

[120] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 3095–3101. IEEE, 2021.

[121] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos. *ArXiv*, abs/2408.10188, 2024.

[122] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 899–908, 2019.

[123] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Fei-Fei Li, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *ArXiv*, abs/2412.14171, 2024.

[124] 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*, 2024.

[125] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

[126] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5684–5693, 2019.

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

[128] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.

[129] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018.

[130] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[131] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *ArXiv*, abs/2406.16852, 2024.

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

[133] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.

[134] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *ArXiv*, abs/2410.02713, 2024.

[135] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025.

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

[137] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024.

[138] Shengjie Zhu, Abhinav Kumar, Masa Hu, and Xiaoming Liu. Tame a wild camera: In-the-wild monocular camera calibration. *Advances in Neural Information Processing Systems*, 36:45137–45149, 2023.

[139] Yiming Zuo and Jia Deng. Ogni-dc: Robust depth completion with optimization-guided neural iterations. In *ECCV*, 2024.

[140] Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. *ICCV*, 2025.