Masked Depth Modeling for Spatial Perception

Bin Tan[1,2], Changjiang Sun[1], Xiage Qin[1], Hanat Adai[1], Zelin Fu[1], Tianxiang Zhou[1], Han Zhang[1], Yinghao Xu[1], Xing Zhu[1], Yujun Shen[1], and Nan Xue[⋆1]

[1] Robbyant, China
[2] Ant Group, China

Abstract. Spatial visual perception is a fundamental requirement in physical-world applications like autonomous driving and robotic manipulation, driven by the need to interact with 3D environments. Capturing pixel-aligned metric depth using RGB-D cameras would be the most viable way, yet it usually faces obstacles posed by hardware limitations and challenging imaging conditions, especially in the presence of specular or texture-less surfaces. In this work, we argue that the inaccuracies from depth sensors can be viewed as "masked" signals that inherently reflect underlying geometric ambiguities. Building on this motivation, we present `MDM`, a depth completion model which leverages visual context to refine depth maps through *masked depth modeling* and incorporates an automated data curation pipeline for scalable training. It is encouraging to see that our model outperforms top-tier RGB-D cameras in terms of both depth precision and pixel coverage. Experimental results on a range of downstream tasks further suggest that `MDM` offers an aligned latent representation across RGB and depth modalities. Code, checkpoints, and $3M$ RGB-depth pairs (including $2M$ real data and $1M$ simulated data): `https://github.com/robbyant/lingbot-depth`.

Keywords: Depth Completion · Masked Depth Modeling · Spatial Perception · RGB-D

1 Introduction

Precise perception of the 3D world is fundamental to any intelligent agent operating in the physical environment—from biological organisms to autonomous vehicles and generalist robots. It provides the essential context for localization ("where am I?") and scene understanding ("what is around me?"). Without robust 3D perception, real-world actions cannot be reliably planned, executed, or verified. Consequently, the pursuit of accurate 3D sensing has become a central pillar of research on physical-grounded artificial intelligence. The optimal approach to achieving this goal remains a subject of debate, with current methods generally falling into three paradigms: (1) classic multi-view visual geometry and recent learning-based adventures, (2) data-driven monocular depth estimation, and (3) active sensor-based depth measurement (*e.g.*, LiDAR, ToF, Structured Light).

⋆ Corresponding author

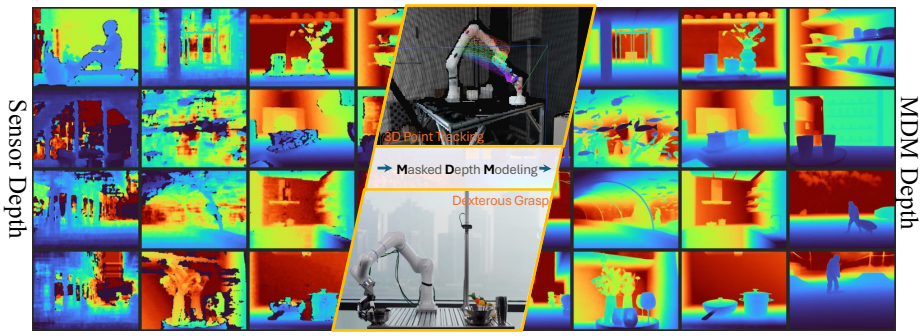


Fig. 1: Enhanced sensor depth powered by our proposed `MDM` pretraining, which leverages naturally missing depth measurements in RGB-D sensors as masks to learn metric-scale, complete, and accurate depth representations. The resulting `MDM` model serves as a powerful spatial perception prior for downstream applications, including 3D point tracking and dexterous grasping.

While the trade-offs of each paradigm have been extensively discussed [11,29,34], the requirements for effective perception came down to three critical criteria: (1) absolute metric scale, (2) pixel-aligned dense geometry, and (3) real-time acquisition without computationally expensive post-processing.

RGB-D cameras distinguish themselves as the only modality capable of satisfying these requirements in real-time. However, their utility is frequently compromised by inherent hardware limitations, particularly the susceptibility of stereo matching algorithms to appearance ambiguities. As illustrated in the left of Fig. 1, even state-of-the-art commercial sensors struggle in challenging scenarios—such as low-texture surfaces, specular reflections, and complex lighting conditions. These failures manifest as severe data corruption and missing values, directly violating the requirement for dense, pixel-aligned geometry.

In this work, we propose a paradigm shift: rather than treating these sensor failures as noise to be discarded, we leverage them as a useful learning signal. Inspired by recent advances in self-supervised masked modeling [6] and joint embedding architectures [1], we introduce **M**asked **D**epth **M**odeling (MDM) to learn dense, pixel-aligned scene geometry. The key innovation of `MDM` lies in treating missing regions ("holes") in raw depth maps as "natural masks", thereby enabling masked depth modeling conditioned on RGB tokens for robust RGB-D representation learning. Because these natural masks arise specifically from geometric and appearance ambiguities (*e.g.*, specular reflections), they present a significantly harder reconstruction challenge than random masking. To address this, our architecture provides the full, unmasked RGB image as conditioning. The model is thus forced to infer missing depth values by reasoning about correlations between the complete RGB context and the remaining valid depth tokens.

By unifying the objectives of monocular depth estimation and depth completion, our `MDM` framework serves as a versatile generalist capable of yielding

metric-scale, pixel-aligned, dense depth maps from arbitrary RGB-D inputs. The transition between these tasks is governed simply by the masking strategy within the Transformer. In the extreme case where all depth tokens are masked, the model functions as a pure Monocular Depth Estimator, forcing the Self-Attention layers to exploit RGB context alone to infer geometry. Conversely, for Depth Completion, we mask only the invalid (sensor-corrupted) tokens, allowing the model to fuse the remaining valid depth readings with visual cues to reconstruct a complete, dense depth prediction.

To support large-scale `MDM` training, we built a scalable data curation pipeline that bridges raw sensor data and reliable supervision. The pipeline includes two parallel streams: a synthetic branch based on self-hosted 3D assets, and a real-world branch using a modular, 3D-printed capture rig compatible with a variety of consumer RGB-D cameras, including active stereo (Intel RealSense, Orbbec Gemini) and passive stereo (ZED) systems. Using this setup, we collected 1M synthetic samples and 2M real captures, each containing synchronized RGB images, raw sensor depth, and stereo pairs. The stereo pairs enable pseudo-depth supervision through a custom stereo matching network adapted from FoundationStereo [34] and trained on synthetic data. We further enrich this corpus with several public RGB-D datasets [3,16,19,20,30,39,41], forming a diverse training set for robust depth completion. Pretraining a ViT-Large on this curated RGB-D corpus with Masked Depth Modeling allows metric geometry to be integrated into semantic tokens via attention, thus improving the sensing quality of RGB-D cameras, as illustrated in the right panel of Fig. 1.

Experimentally, we first validate `MDM` on depth completion benchmarks, where it achieves competitive performance. Moreover, it serves as a stronger monocular depth prior for FoundationStereo [34] than DepthAnythingV2 [38], improving stereo matching performance. Our pretrained model demonstrates strong zero-shot generalization to 3D tracking tasks. When deployed as a drop-in metric depth estimator in SpatialTrackerV2 [36], replacing the VGGT [27] frontend, it leads to improved motion understanding and higher computational efficiency in real world. Finally, we demonstrate practical robotic utility through dexterous grasping. Our approach leverages robust depth predictions and learned representations from `MDM` to enable open-world grasping without requiring object CAD models or perfect simulator states. We successfully grasp challenging objects that typically confound conventional depth sensors, including transparent glassware and highly reflective surfaces.

2 Related Works

2.1 Metric Depth Estimation

Estimating depth at absolute metric scale is essential for physical-world applications such as robotic manipulation and autonomous driving. Foundation monocular depth models [2,8,17,28,29,37,38,40] have dramatically improved relative depth quality, yet they are inherently scale-ambiguous and struggle to produce stable metric predictions in real-world conditions. A complementary line of work

conditions depth estimation on an imperfect metric prior from an active sensor, lifting the scale ambiguity while relying on visual features to fill corrupted or missing regions. Methods in this family differ primarily in *where* the depth prior is fused with RGB features. OMNI-DC [44] operates in a gradient domain: it predicts depth gradients jointly from RGB and sparse depth inputs and recovers a dense depth map via a differentiable depth-integration layer, achieving robustness across highly irregular sparsity patterns. PromptDA [11] fuses on the *decoder* side—a lightweight convolutional network encodes sparse metric depth and injects the resulting features into the decoder of DepthAnythingV2 [38], steering its relative predictions toward metric scale at up to 4K resolution. Conversely, PriorDA [31] fuses on the *encoder* side, blending depth conditions with the RGB input through a shallow convolutional layer *before* the ViT patch-embedding step, allowing the transformer to jointly attend over both modalities from the very first layer. CDMs [12] adopts a dual-branch ViT design that tokenises RGB and depth independently and merges the two token streams before decoding, enabling richer cross-modal interaction for manipulation-aware metric reconstruction. Our MDM differs from all of the above in training paradigm: rather than designing task-specific fusion modules, we train a single ViT on joint RGB–depth tokens via Masked Depth Modeling, letting attention naturally integrate metric geometry into visual representations as a pretraining objective.

2.2 Mask Modeling

Masked Autoencoders (MAE) [6] established masked patch prediction as a scalable pre-training paradigm for Vision Transformers: a high fraction of image patches is randomly removed, and the model is trained to reconstruct the missing raw pixels. Despite its simplicity, MAE learns representations that transfer competitively to a wide range of downstream vision tasks. I-JEPA [1] proposes a Joint-Embedding Predictive Architecture (JEPA) that predicts abstract *latent* representations of masked target blocks from a context encoder, rather than regressing raw pixels. By operating in a learned feature space, JEPA avoids the need to model low-level texture and yields richer semantic representations. CroCo [32, 33] extends masked modeling to 3D vision by masking one view of a stereo pair and reconstructing it from the other, demonstrating that cross-view completion is a powerful pre-text task for depth estimation and stereo matching. Our work draws inspiration from these masked modeling paradigms, but repurposes the masking signal for RGB-D perception. By masking depth tokens and conditioning reconstruction on the full RGB context, we force the model to learn tight cross-modal dependencies between appearance and geometry, yielding robust RGB-D representations for downstream tasks.

3 Masked Depth Modeling

Masked depth modeling follows the general paradigm of masked image modeling [6] within an encoder–decoder framework, but shifts the learning objective from

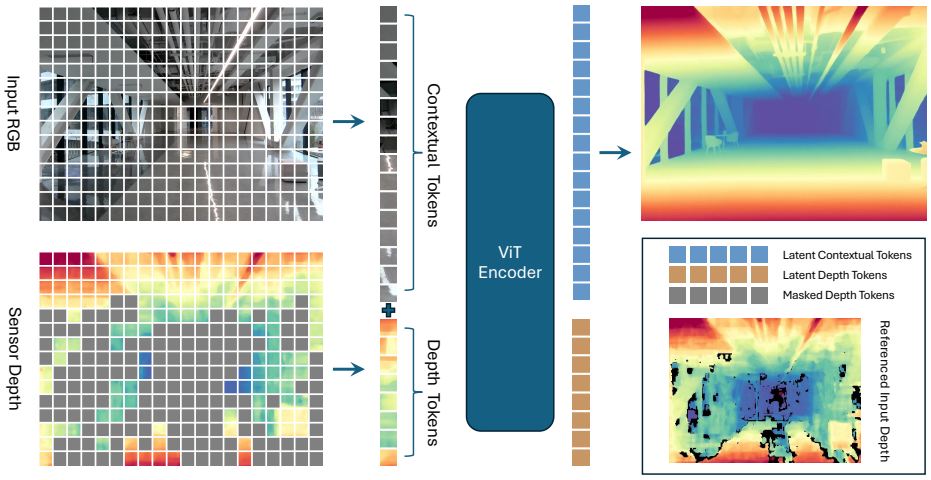


Fig. 2: Illustration of the proposed Masked Depth Modeling framework. Depth tokens corresponding to missing sensor measurements are masked, and a ViT encoder learns a joint embedding from the contextual tokens (i.e., the RGB frame) and the remaining unmasked depth tokens. In the decoder stage, latent depth tokens are discarded, and a ConvStack decoder reconstructs the full depth map from latent contextual tokens. We put an unmasked depth map in the bottom-right as the reference.

appearance reconstruction to depth map prediction by operating on RGB-D inputs. We adopt a standard Vision Transformer (ViT) architecture [4], using ViT-Large as the encoder backbone. For depth reconstruction from patch tokens, instead of the shallow Transformer decoder used in vanilla MAE [6], we employ a ConvStack decoder adopted from MoGe [28, 29], which is better suited for dense geometric prediction. Figure 2 provides an overview of the masked depth modeling architecture.

3.1 Separated Patch Embedding for RGB-D Inputs

Patch Embedding Layers. We apply separate patch embedding layers to the two input modalities: the 3-channel RGB image and the single-channel depth map. Each modality is independently projected into a sequence of patch tokens, yielding RGB tokens and depth tokens that are spatially aligned on the same 2D grid. This separated patch embedding design enables the self-attention layers in the ViT encoder to learn joint representations that integrate appearance context from RGB images with geometric cues from depth measurements. In particular, the model can exploit complementary information from rich visual context and fundamental geometric priors such as near–far relationships, coplanarity, and spatial continuity. In our implementation, we set the patch size to 14 for both patch embedding layers, following DINOv2 [15]. Without loss of generality, we assume that the RGB frame and the depth map share the same spatial resolution (H, W), where both H and W are divisible by 14. The number of tokens per

modality is therefore $N = HW/14^2$. We denote the i-th RGB token as $\mathbf{c}_i \in \mathbb{R}^n$ and the i-th depth token as $\mathbf{d}_i \in \mathbb{R}^n$, where n is the token embedding dimension.

Positional Embeddings. RGB-D sequences require encoding (1) 2D spatial location and (2) modality identity (RGB vs. depth) for each token. We use two embeddings: (1) shared learnable 2D spatial embedding for both modalities, and (2) modality embedding (1 for RGB, 2 for depth). The final positional encoding is the sum of spatial and modality embeddings, added to each token before the attention blocks.

3.2 Joint Embedding of RGB and Unmasked Depth for Masked Depth Prediction

Depth sensors are inherently sensitive to appearance-related disturbances, and it is common for RGB-D cameras to produce depth maps with missing or invalid measurements. Importantly, such missing depth values often correlate with challenging visual conditions in the scene, including surface material properties, lighting variations, and reflective or textureless regions. Rather than treating these missing measurements as noise, we view them as an opportunity to learn a joint representation from the complementary sources: valid depth observations and the always-available RGB appearance context. Accordingly, we leverage the masking patterns naturally induced by depth sensors and train the model to learn a joint embedding of RGB tokens and unmasked depth tokens, enabling robust depth reasoning under incomplete observations.

Masking from Missing Depth Measurements. Our masking strategy is grounded in the inductive bias introduced by missing depth measurements. In practice, however, many RGB-D samples are well conditioned and contain few or no missing depth values. Such samples are nonetheless valuable for learning joint embeddings, as they provide abundant paired appearance–geometry observations and should not be discarded. Accordingly, we aim to incorporate as many RGB-D samples as possible to support large-scale training of Vision Transformers. Another practical consideration arises from patch-based processing: a single patch may contain a mixture of valid and invalid depth values. To address this ambiguity, we define the masking decision at the patch level based on the validity statistics within each patch, ensuring a consistent and well-defined masking rule for depth tokens.

Based on the above discussion, a depth patch (token) whose values are entirely missing is always masked. For patches containing a mixture of valid and invalid depth values, we assign a higher probability (*e.g.*, 0.75 in our training) of being masked. If the number of masked tokens from these two cases is insufficient to meet the target masking ratio, we randomly sample additional fully valid depth tokens to complete the masking set. Such a masking strategy allows imperfect yet informative depth tokens to remain unmasked, enabling meaningful interaction with contextual RGB tokens. The overall masking ratio is in the range of 60%-90% for depth maps.

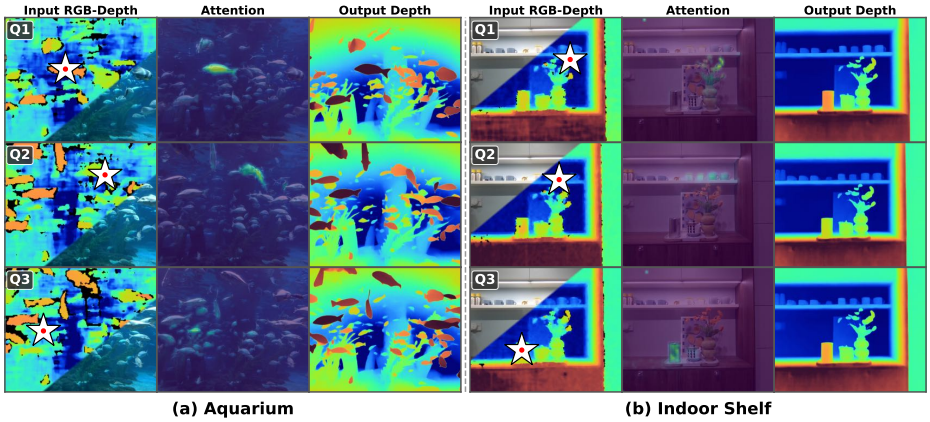


Fig. 3: Multi-query depth-to-RGB attention visualization. For two scenes:(a) densely packed aquarium and (b) indoor shelf with diverse materials, we select three depth query patches (Q1–Q3) and visualize their attention over RGB tokens. Each row shows masked depth (query marked by $\star$), attention overlay on RGB, and refined depth. Different queries attend to distinct, spatially corresponding regions, confirming fine-grained cross-modal geometric-appearance associations.

RGB-D Tokens for Vision Transformers. After applying the masking strategy, the full set of RGB tokens and the unmasked depth tokens are concatenated to form RGB-D tokens, which are fed into a ViT-Large encoder with 24 self-attention blocks. In addition to the RGB-D tokens, a `[cls]` token is retained to capture global context across modalities. Different from conventional designs that aggregate tokens from multiple intermediate layers (e.g., layers 6, 12, 18, and 24) for dense prediction, as adopted in DepthAnythingV2 [38] and MoGe [28,29], we only retain the output tokens from the final encoder layer for subsequent processing.

In vanilla MAE, pixel-wise RGB reconstruction is performed using a shallow Transformer decoder. Each patch token is decoded into $3 \times$ `patch_size` $\times$ `patch_size` values, which are then reshaped into the pixel domain. This design is suitable for RGB-only masking, where masked tokens share homogeneous latent representations within the decoder. In contrast, our masking strategy is applied exclusively to depth tokens, while the RGB tokens remain fully observable and provide complete contextual information. This setting allows depth prediction at each spatial location to be conditioned on rich appearance context. Consequently, we adopt a convolutional decoder architecture, termed *ConvStack*, which is more appropriate for dense geometric reconstruction.

ConvStack Decoder. After encoding, we discard depth tokens and retain contextual tokens as spatial representations. The `[cls]` token is broadcast and added to each contextual token to inject global context. These tokens serve as queries to a hierarchical decoder from MoGe [28, 29]. The decoder has a shared neck and task-specific heads. Starting from (h, w), the neck upsamples features via

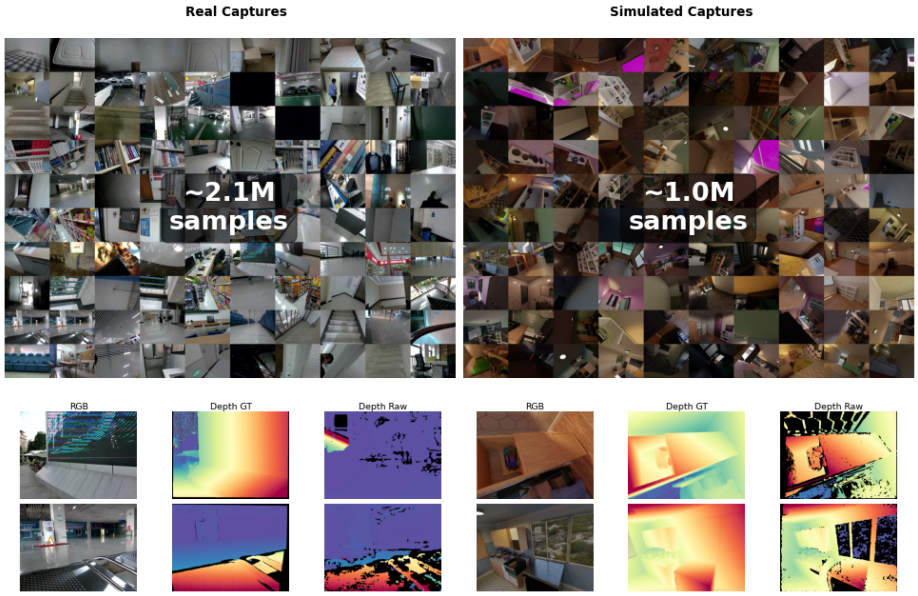


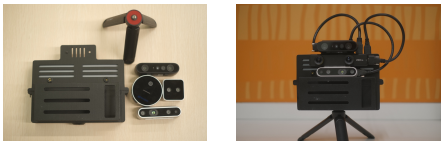
Fig. 4: Our data curation pipelines. Samples from a total of 2.1M real-captured samples plus 1.0M simulated captures are gather in the top row. In the bottom row, we show the RGB-D inputs and the GT depth maps accordingly.

residual blocks and transposed convolutions (kernel 2, stride 2), doubling resolution each stage to $(16h, 16w)$. Multi-scale features are shared across heads for efficient reuse, while each head produces its own output. The predicted depth is finally upsampled to input resolution.

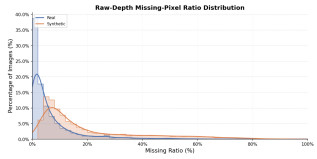
Attention Visualization. We visualize depth-to-RGB attention from the final encoder layer by projecting attention weights onto the image as heatmaps (Fig. 3). Across diverse scenarios (packed aquarium, indoor shelf), depth tokens consistently attend to spatially localized RGB regions matching their query locations. This confirms fine-grained, position-aware geometric-appearance learning rather than trivial patterns.

3.3 Training

We use a 24-layer ViT-L/14 [15] encoder initialized with DINOv2 weights. The decoder with shared neck and task-specific heads is randomly initialized. Training runs for 250K iterations with batch size 1,024 (128 GPUs). We use gradient clipping (norm 1.0) and BF16 precision, taking about 7.5 days. Depth is supervised with L1 loss on valid pixels only. See supplement for details.



(a) RGB-D capture system. Hardware components (left) and assembled system (right).



(b) Raw depth missing ratio distributions for synthetic and real datasets.

Fig. 5: Data curation. (a) Our scalable capture hardware. (b) Raw depth missing ratio distributions.

Table 1: Distribution of our real-world RGB-D captures by scene category. We design a diverse collection protocol spanning residential, commercial, public, and specialized spaces to ensure broad coverage of indoor environments.

Residential Space		**Work / Study Space**		**Commercial / Service**	
$< 50\,\mathrm{m}^2$	10.16%	Office	3.39%	Retail Store	3.39%
50–$100\,\mathrm{m}^2$	10.16%	Meeting Room	3.39%	Restaurant / Café	3.39%
$> 100\,\mathrm{m}^2$	10.16%	Classroom	3.39%	Hotel / Guestroom	1.70%
		Library	1.70%	Gym	3.39%
		Workshop	1.70%	Lobby / Corridor	1.70%
		Laboratory	1.70%		
Public Space		**Special-Function**		**Outdoor Scene**	
Hospital / Clinic	3.39%	Parking Garage	3.39%	Outdoor Environment	10.16%
School / Hall	3.39%	Machinery Room	3.39%		
Airport / Station	3.39%	Elevator	3.39%		
Museum	3.39%	Corridor / Passage	3.39%		
		Storage / Warehouse	3.39%		

4 Data Curation Pipelines

We curate RGB-D data with realistic missing patterns via two streams: synthetic data from 3D assets (`MDM-S`) and real-world data from commercial depth cameras (`MDM-R`) (raw depth missing ratio distributions in Fig. 5b).

Synthetic Data Pipeline. Unlike approaches rendering idealized depth, we simulate real RGB-D camera imaging to generate realistic depth with natural imperfections. Using 3D assets in Blender, we render RGB, perfect depth, and stereo pairs with speckle patterns. RGB is rendered from the left camera for pixel-wise alignment with stereo depth. Stereo images are processed with SGM [7] to produce sensor-like depth mimicking real artifacts. We sample baseline from $[0.05, 0.2]$m and focal length from $[16, 28]$mm for diverse geometries, rendering 1M samples from 442 indoor scenes. Examples in supplement.

Scalable RGB-D Capture Systems. We build a scalable capture system (Fig. 5a) with 3D-printed fixtures mounting commercial RGB-D cameras and a touchscreen PC managing data streams. We deploy multiple devices across scenes in Tab. 1. Since real captures lack perfect depth, we compute stereo disparity from

Table 2: Depth completion results. (a) Block-wise depth masking with four difficulty levels. (b) Sparse depth inputs.

(a) Protocol 1: Block-wise Depth Masking.

	Easy		Medium		Hard		Extreme		Easy		Medium		Hard		Extreme	
	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓	RMSE↓	REL↓
	iBims-1 [10]								NYUv2 [23]							
CDM [12]	0.958	0.271	1.476	0.462	2.021	0.681	2.741	0.978	0.372	0.098	0.558	0.160	0.711	0.209	1.310	0.446
OMNI-DC [44]	0.476	0.084	0.929	0.232	1.382	0.342	2.053	0.555	0.234	0.055	0.308	0.079	0.395	0.103	0.643	0.176
OMNI-DC-DA [44]	0.602	0.068	0.771	0.155	1.266	0.302	1.937	0.538	0.343	0.053	0.385	0.070	0.418	0.088	0.456	0.116
PriorDA [31]	0.409	0.095	0.433	0.093	0.520	0.094	0.845	0.150	0.289	0.093	0.276	0.080	0.267	0.072	0.309	0.076
PromptDA [11]	0.298	0.055	0.337	0.065	0.428	0.083	0.607	0.129	0.209	0.051	0.219	0.053	0.238	0.057	0.324	0.074
Ours	**0.186**	**0.016**	**0.213**	**0.025**	**0.250**	**0.040**	**0.303**	**0.063**	**0.072**	**0.011**	**0.108**	**0.017**	**0.124**	**0.021**	**0.167**	**0.029**
	DIODE-Indoor [25]								DIODE-Outdoor [25]							
CDM [12]	0.711	0.283	0.956	0.385	1.204	0.480	1.889	0.724	11.630	0.252	11.772	0.348	12.185	0.448	13.074	0.579
OMNI-DC [44]	0.563	0.074	0.895	0.121	1.108	0.153	1.775	0.271	2.214	0.041	2.806	0.067	3.701	0.103	6.239	0.204
OMNI-DC-DA [44]	0.710	0.065	0.991	0.097	1.222	0.121	1.592	0.211	2.229	0.053	3.703	0.070	5.352	0.110	5.988	0.173
PriorDA [31]	0.370	0.073	0.354	0.064	0.369	0.062	0.665	0.083	2.734	0.090	2.693	0.085	2.909	0.088	5.114	0.136
PromptDA [11]	0.250	0.049	0.291	0.054	0.320	0.060	0.465	0.083	2.587	0.070	2.807	0.087	3.026	0.100	4.313	0.156
Ours	**0.077**	**0.007**	**0.148**	**0.015**	**0.150**	**0.016**	**0.219**	**0.022**	**2.019**	**0.036**	**2.436**	**0.047**	**2.654**	**0.054**	**3.913**	**0.085**

(b) Protocol 2: Sparse Depth Inputs.

Method (depth pattern→)	VOID [35] sparse-500	VOID [35] sparse-1500	KITTI DC [5] lidar-wise	ETH3D [22] sparse-SfM	ETH3D [22] lidar-wise	Avg. Rank
OMNI-DC-DA [44]	0.551	0.554	**1.310**	0.605	**0.101**	3.00
Marigold-DC [26]	0.535	0.551	1.465	2.330	0.225	4.20
PriorDA [31]	0.524	0.543	1.421	0.360	0.107	2.60
Any2Full [43]	0.537	0.561	1.404	0.562	0.168	3.60
Ours	**0.487**	**0.516**	1.412	**0.152**	0.106	**1.60**

IR pairs following FoundationStereo [34], applying left-right consistency checks for quality. We collect 2M real captures with diverse scenes for `MDM`.

4.1 Training Data Summary

Beyond our curated 3.2M samples (`MDM-S` and `MDM-R`), we incorporate open-source datasets [3, 16, 19, 20, 30, 39, 41] to form 10M training samples total. For open-source synthetic datasets with complete depth, we randomly mask 60%–90% of tokens. For real-world open-source datasets, which are more complete than our curated data, masking is also primarily random. More details are in the supplementary material.

5 Experiments

We validate the effectiveness of `MDM` pretraining through experiments. First, we evaluate `MDM` on depth completion (Sec. 5.1), which aligns closely with our pretraining objective. Second, we show that using our MDM-pretrained model improves FoundationStereo over its original DepthAnythingV2 [38] backbone (Sec. 5.2). Finally, Sec. 6 presents three downstream applications: video depth completion, 3D point tracking, and robotic dexterous grasping.

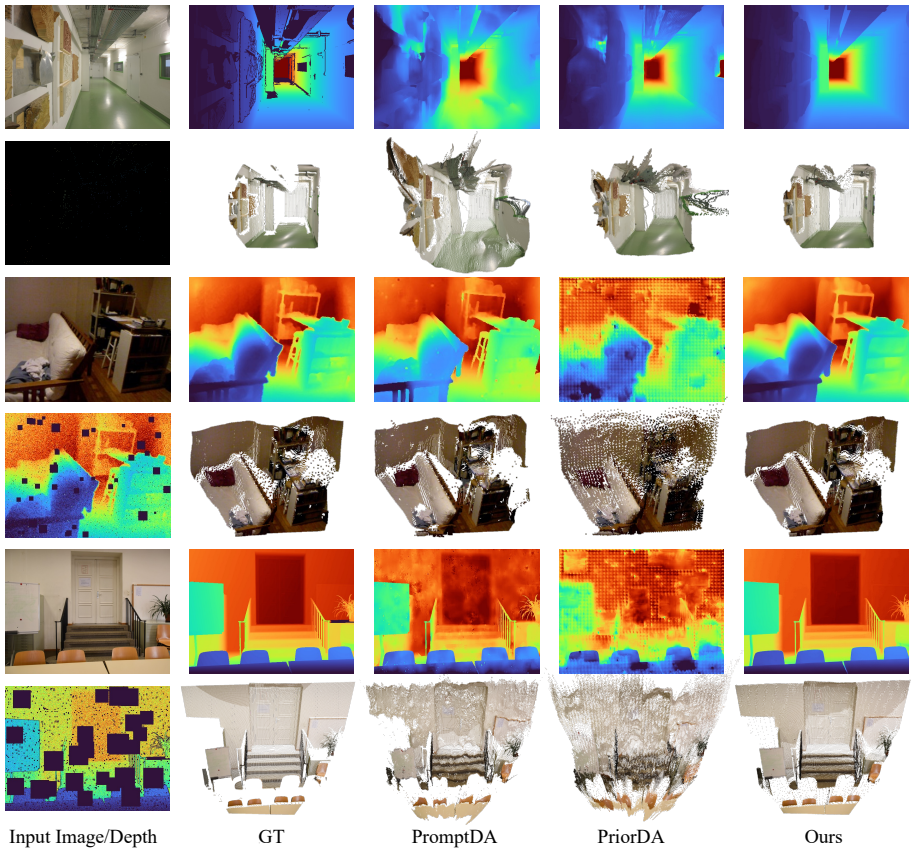


Fig. 6: Comparison of depth completion. We show the RGB input, sparse/masked depth input, and predictions from PromptDA, PriorDA, and ours.

5.1 Depth Completion

As `MDM` pretraining is naturally aligned with the depth completion task, we evaluate our `MDM` model against state-of-the-art approaches, including CDM [12], OMNI-DC [44], PromptDA [11], and PriorDA [31]. The evaluation checkpoints for these methods are obtained from their official repositories. We follow two evaluation protocols, (1) Block-wise Depth Masking and (2) Sparse and Dense Depth Inputs.

Protocol 1: Block-wise Depth Masking. We generate incomplete depth by randomly masking spatial blocks from ground-truth depth, simulating dropout in consumer depth cameras. To model low-quality measurements, we add Gaussian noise and shot-noise perturbations based on Kinect noise models [14]. Each dataset has four difficulty levels: *easy*, *medium*, *hard*, and *extreme*. We evaluate on iBims-1 [10], NYUv2 [23], and DIODE [25].

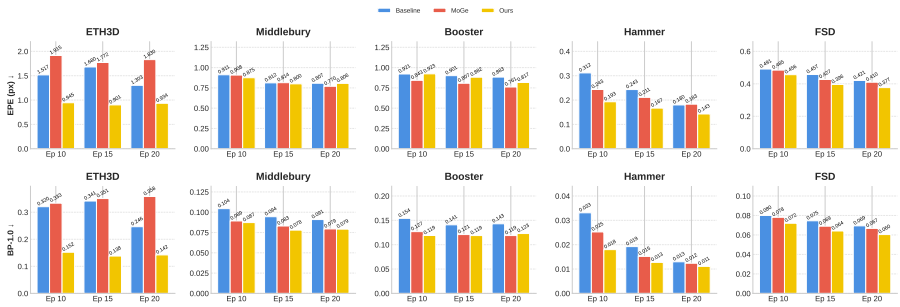


Fig. 7: Epoch-wise comparison of FoundationStereo with different depth priors. EPE (top) and BP-1.0 (bottom) at epochs 10, 15, and 20 on five benchmarks.

Protocol 2: Sparse Depth Inputs. SfM/SLAM techniques are commonly used to recover camera poses and sparse scene geometry when depth cameras are unavailable. Recovering complete depth from sparse points is valuable for downstream applications but challenging due to extreme sparsity, making this protocol more difficult than Protocol 1. We evaluate our model on the ETH3D [22], IBims-1 [10], KITTI-DC [5], VOID [35] with sparse observations.

Results. As shown in Tab. 2a, our method outperforms all competitors across all difficulty levels on every dataset under Protocol 1. On indoor benchmarks (iBims, NYUv2, DIODE-Indoor), `MDM` reduces RMSE by over 40% relative to the best competitor (PromptDA) even in the *extreme* setting, demonstrating resilience to heavy masking and noise. On DIODE-Outdoor with larger depth ranges, our method achieves the lowest errors across all levels. Under Protocol 2 (Tab. 2b), with only sparse SfM depth as input, `MDM` achieves state-of-the-art results on ETH-SfM compared to the best baseline. Qualitative comparisons (Fig. 6) show `MDM` produces sharper boundaries and more coherent structures than competitors, especially in missing or sparse regions. These results confirm that `MDM` pretraining equips the model with strong depth priors that generalize across diverse corruption and sparsity patterns.

5.2 FoundationStereo with MDM Pretraining

We use our pretrained `MDM` as a monocular depth prior in FoundationStereo [34]. Following their meta-architecture, we adapt `MDM`'s RGB-only branch as depth priors while keeping other components unchanged. For comparison, we train three variants using the same hyperparameters: (1) FoundationStereo with our pretrained `MDM`, (2) with MoGe [29], and (3) vanilla FoundationStereo with DepthanythingV2 [38]. All models are trained for 15 epochs on FSD [34] without iterative data curation to save compute.

Table 3: Camera pose estimation on TUM-RGBD [24].

Depth Type	ATE↓	RPE_t↓	RPE_r↓
Raw	0.131	**0.038**	0.844
Ours	**0.118**	0.039	**0.791**

Evaluation. We evaluate on ETH3D [22], Middlebury [21], Booster [18], HAMMER [9], and FSD [34], reporting EPE (end-point error) and BP-1.0 (percentage of pixels with disparity error >1.0).

Results. Fig. 7 shows performance at epochs 10, 15, and 20. Key observations: (1) **Faster convergence.** At epoch 10, our variant outperforms vanilla on most benchmarks. (2) **Final performance.** At epoch 20, our variant achieves best or comparable results across all benchmarks, confirming `MDM` pretraining provides effective initialization for stereo matching.

6 Extensions and Applications

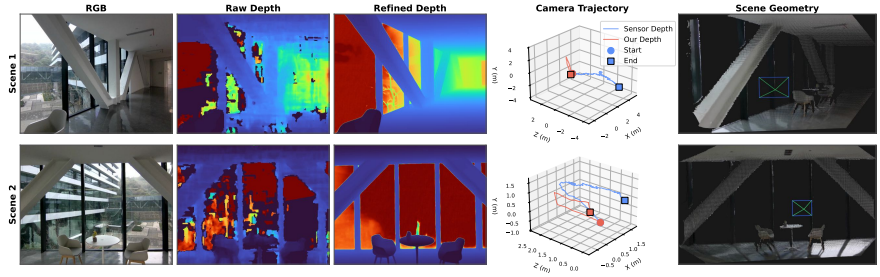
We showcase applications built on `MDM` using an Orbbec Gemini-335 RGB-D camera. These applications demonstrate why RGB-D cameras are important for spatial perception and their potential to shape future vision systems.

6.1 Online 3D Point Tracking

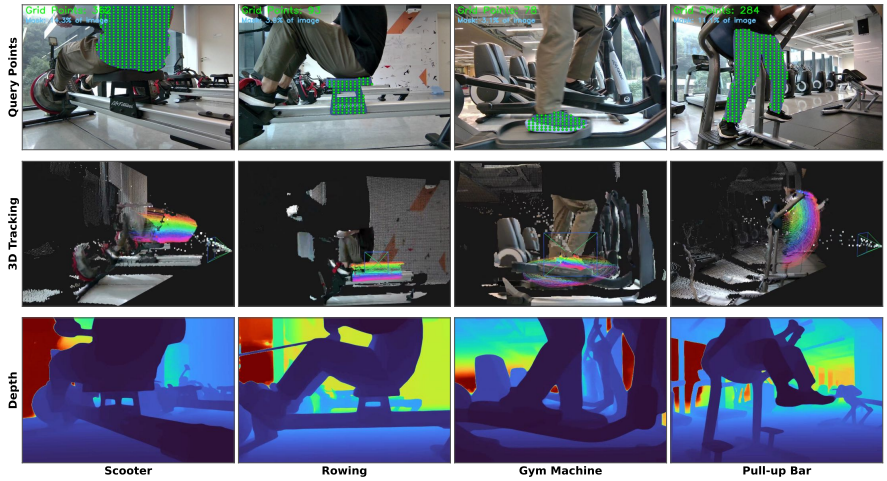
Leveraging strong depth completion performance of our `MDM`, we use SpatialTrackerV2 [36] for camera pose estimation and 3D point tracking. We adopt its online version (without VGGT [27] frontend) as an RGB-D tracking baseline. Unlike the original implementation assuming known camera extrinsics, we initialize frame-wise extrinsics with SE(3) identity and rely on Bundle Adjustment (BA) on tracked VO points for camera motion. We use no global BA for efficiency and no fine-tuning of SpatialTrackerV2.

Camera Tracking. In Tab. 3, we evaluate camera poses calculated from tracked points on the TUM-RGBD dynamic scenes [24]. The estimated camera poses are effectively improved by using our refined depth. Fig. 8a shows camera tracking on two indoor scenes with extensive glass surfaces where raw depth sensors fail. Using our refined depth, SpatialTrackerV2 produces smoother and more accurate trajectories compared to raw sensor depth, which suffers from severe drift due to missing regions.

Object Motion Tracking. Fig. 8b shows dynamic 3D point tracking on four scenarios with moving objects. Given query points on objects of interest, SpatialTrackerV2 successfully tracks their 3D trajectories using our refined depth, while



(a) Camera tracking and scene reconstruction. Left to right: RGB input, raw sensor depth, our refined depth, camera trajectories (sensor: blue, ours: red), and reconstructed geometry.



(b) Dynamic 3D point tracking. Top: query points on target object. Middle: tracked trajectories (rainbow-colored by time). Bottom: depth maps of our MDM.

Fig. 8: Online 3D point tracking with SpatialTrackerV2 using our refined depth.

tracking fails with raw depth due to noise and missing measurements. Rainbow-colored trails reveal coherent motion patterns, confirming our depth maintains geometric accuracy for downstream tracking.

6.2 Grasp Pose Generation in Real-World Robotics

We apply MDM to real-world dexterous grasping, where accurate depth is critical for grasp pose generation. Our system uses a Rokae XMate-SR5 arm with X Hand-1 dexterous hand and an Orbbec Gemini 335 RGB-D camera. Given RGB-D input, we convert depth to point cloud and predict $N\times22$ hand poses using a diffusion policy conditioned on DINOv2 RGB features and Point Transformer point cloud features, following DP3 [42]. The model is trained on the HOI4D [13], where human hand-object interactions are retargeted to the dexterous hand via 3D keypoint correspondence.

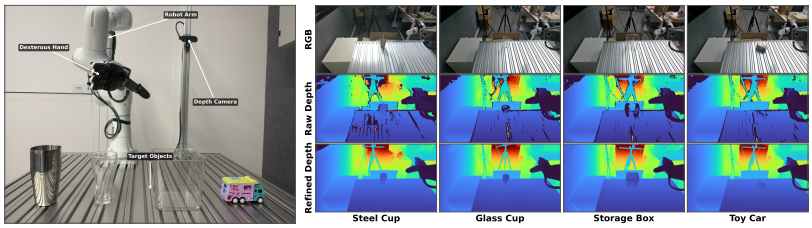


Fig. 9: Qualitative grasping results. Left: hardware setup. Right: RGB, raw depth, and refined depth for four objects.

Table 4: Grasping success rates with our refined depth vs. raw sensor depth on four challenging objects.

	Total Times	with `MDM`	with raw depth
Stainless steel cup	20	17	13
Transparent cup	20	16	12
Toy Car	20	16	9
Transparent storage box	20	10	N/A

Results. We evaluate grasping on four challenging objects, including transparent and reflective items where raw depth sensors fail. As shown in Tab. 4, `MDM` consistently improves success rates over raw sensor depth. Notably, the transparent storage box is ungraspable with raw depth (N/A) due to severe corruption, whereas our model achieves 50% success by producing plausible depth estimates despite occasional inaccuracies on highly transparent surfaces. This shows that `MDM`'s improved depth enables reliable robotic manipulation.

7 Conclusion

In this paper, we investigate missing depth measurements in RGB-D cameras. Rather than treating these as pure sensor failures, we leverage them as natural masks reflecting appearance ambiguities in depth imaging, proposing Masked Depth Modeling (MDM). Using 3 million self-curated RGB-D samples with open-source datasets, we pretrain a Vision Transformer to learn joint representations of appearance and depth. The resulting model, `MDM`, demonstrates strong depth completion performance and serves as a powerful monocular depth prior for FoundationStereo. Beyond static tasks, we validate its effectiveness in downstream applications including 3D point tracking and dexterous grasping, demonstrating `MDM`'s broad applicability.

Acknowledgements

This work was supported by Ant Group Postdoctoral Programme. We gratefully acknowledge the optical testing team at Orbbec Inc. for their professional support.

References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M.G., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15619–15629 (2023)
2. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: Int. Conf. Learn. Represent. (2025)
3. Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., Shulman, E.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In: Adv. Neural Inform. Process. Syst. (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3354–3361 (2012)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15979–15988 (2022)
7. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. **30**(2), 328–341 (2008)
8. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Trans. Pattern Anal. Mach. Intell. **46**(12), 10579–10596 (2024)
9. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., Leonardis, A., Busam, B.: Is my depth ground-truth good enough? HAMMER – highly accurate multi-modal dataset for dense 3d scene regression. CoRR **abs/2205.04565** (2022)
10. Koch, T., Liebel, L., Körner, M., Fraundorfer, F.: Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. Comput. Vis. Image Underst. **191**, 102877 (2020)
11. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17070–17080 (2025)
12. Liu, M., Zhu, Z., Han, X., Hu, P., Lin, H., Li, X., Chen, J., Xu, J., Yang, Y., Lin, Y., Li, X., Yu, Y., Zhang, W., Kong, T., Kang, B.: Manipulation as in simulation: Enabling accurate geometry perception in robots. In: Int. Conf. Learn. Represent. (2026)
13. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: HOI4D: A 4d egocentric dataset for category-level human-object interaction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21013–21022 (2022)
14. Nguyen, C.V., Izadi, S., Lovell, D.R.: Modeling kinect sensor noise for improved 3d reconstruction and tracking. In: International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT). pp. 524–530 (2012)

15. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Trans. Mach. Learn. Res. **2024** (2024)
16. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O.M., Newcombe, R.A., Ren, C.Y.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: Int. Conf. Comput. Vis. pp. 20076–20086 (2023)
17. Piccinelli, L., Yang, Y., Sakaridis, C., Segù, M., Li, S., Gool, L.V., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10106–10116 (2024)
18. Ramirez, P.Z., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Stefano, L.D.: Open challenges in deep stereo: The booster dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21168–21178 (2022)
19. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.Á., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Int. Conf. Comput. Vis. pp. 10892–10902 (2021)
20. Sajjan, S.S., Moore, M., Pan, M., Nagaraja, G., Lee, J., Zeng, A., Song, S.: Cleargrasp: 3d shape estimation of transparent objects for manipulation. In: IEEE International Conference on Robotics and Automation (ICRA). pp. 3634–3642 (2020)
21. Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P.: High-resolution stereo datasets with subpixel-accurate ground truth. In: Pattern Recognition: 36th German Conference. pp. 31–42 (2014)
22. Schöps, T., Schönberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2538–2547 (2017)
23. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Eur. Conf. Comput. Vis. vol. 7576, pp. 746–760 (2012)
24. Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of rgb-d slam systems. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 573–580 (2012)
25. Vasiljevic, I., Kolkin, N.I., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A dense indoor and outdoor depth dataset. CoRR **abs/1908.00463** (2019)
26. Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K., Obukhov, A.: Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In: Int. Conf. Comput. Vis. pp. 5359–5370 (2024)
27. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
28. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5261–5271 (2025)
29. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: Adv. Neural Inform. Process. Syst. (2025)

30. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.A.: Tartanair: A dataset to push the limits of visual SLAM. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4909–4916 (2020)
31. Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H., Zhao, Z.: Depth anything with any prior. arXiv preprint arXiv:2505.10565 (2025)
32. Weinzaepfel, P., Leroy, V., Lucas, T., Brégier, R., Cabon, Y., Arora, V., Antsfeld, L., Chidlovskii, B., Csurka, G., Revaud, J.: Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion. In: Adv. Neural Inform. Process. Syst. (2022)
33. Weinzaepfel, P., Lucas, T., Leroy, V., Cabon, Y., Arora, V., Brégier, R., Csurka, G., Antsfeld, L., Chidlovskii, B., Revaud, J.: Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In: Int. Conf. Comput. Vis. pp. 17923–17934 (2023)
34. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5249–5260 (2025)
35. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. IEEE Robotics and Automation Letters **5**, 1899–1906 (2019)
36. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. In: Int. Conf. Comput. Vis. (2025)
37. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. (2024)
38. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything V2. In: Adv. Neural Inform. Process. Syst. (2024)
39. Yeshwanth, C., Liu, Y., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Int. Conf. Comput. Vis. pp. 12–22 (2023)
40. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from A single image. In: Int. Conf. Comput. Vis. pp. 9009–9019 (2023)
41. Zamir, A.R., Sax, A., Shen, W.B., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3712–3722 (2018)
42. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In: Robotics: Science and Systems (RSS) (2024)
43. Zhou, Z., Liu, R., Liu, T., Zuo, W., Wang, S., Hong, Z., Zhang, D.: Any to full: Prompting depth anything for depth completion in one stage. ArXiv **abs/2603.05711** (2026)
44. Zuo, Y., Yang, W., Ma, Z., Deng, J.: OMNI-DC: Highly robust depth completion with multiresolution depth integration. In: Int. Conf. Comput. Vis. (2025)