DepthMaster: Unified Monocular Depth Estimation for Perspective and Panoramic Images

Pengfei Wang, Shihao Wang, Liyi Chen,
Zhiyuan Ma, Guowen Zhang, Lei Zhang[†]
The Hong Kong Polytechnic University
pengfei.wang@connect.polyu.hk cslzhang@comp.polyu.edu.hk
[†]Corresponding author

Abstract

While monocular depth estimation has achieved significant progress, achieving generalized metric depth estimation for both narrow field-of-view (FoV) perspectives and 360° panoramas remains an unsolved challenge. Existing methods are often tailored to specific camera types and struggle to produce accurate metric depth that generalizes across diverse settings. This limitation stems from two key challenges: the inherent geometric discrepancy between perspective and panoramic cameras, and the scarcity of panoramic training data with metric annotations. In this work, we introduce **DepthMaster**, a unified metric depth estimation framework. Rather than employing specialized networks to learn spherical distortions, we reformulate the problem by decomposing panoramic images into overlapping perspective patches. Crucially, distinct from prior projection-based methods that rely on ad-hoc architectural modifications to handle boundaries, we introduce a novel Correspondence Consistency Loss (CCL) and inject virtual projection cameras as geometric priors, allowing us to seamlessly stitch the patches while avoiding specialized operators and keeping the backbone largely compatible with standard Transformer designs. This strategy also resolves the geometric differences by unifying all inputs into a canonical perspective representation, and effectively circumvents data scarcity by directly unlocking powerful metric priors from vast perspective datasets. Trained on a mixed dataset that contains only one panorama dataset, DepthMaster achieves state-of-the-art zero-shot performance on 13 diverse datasets, outperforming **not only universal methods but also leading specialist models in both perspective and panoramic domains**. Project page: `https://polyu-vclab.github.io/DepthMaster-page`

1 Introduction

Monocular depth estimation (MDE), particularly metric depth estimation, is a fundamental task in computer vision, which has wide applications in 3D scene understanding and generation [24, 43, 69, 10, 68, 11, 44, 91, 92], spatial reasoning [71, 33, 76, 55, 35], autonomous navigation, augmented reality, and robotics [94, 23, 19, 42]. In recent years, MDE has achieved remarkable progress, largely driven by deep learning models trained on large-scale, diverse, and high-quality data [89, 22, 85, 86, 52, 72]. State-of-the-art (SOTA) methods [90, 50, 85, 86, 6, 73] have demonstrated impressive zero-shot generalization performance, capable of recovering high-fidelity metric depth from diverse in-the-wild perspective images.

However, the success of these SOTA models is predominantly confined to standard pinhole cameras, and their performance degrades dramatically on 360° panoramic images due to severe equirectangular projection (ERP) distortions. While specialist models have been developed for specific domains

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

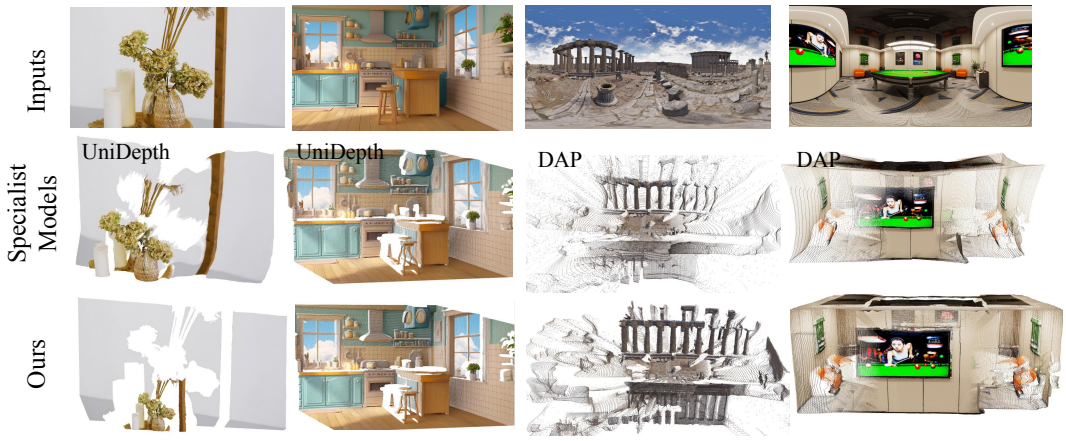


Figure 1: **DepthMaster enables high-fidelity zero-shot metric depth estimation.** We visualize the 3D point clouds reconstructed from diverse inputs. While specialist models—such as UniDepth [50] for perspective images and Depth Any Panoramas (DAP) [41] for panoramas—often struggle with global geometric structures or local fine details, our unified DepthMaster produces superior 3D structures and sharper details across both camera types. We provide more interactive visualizations on our project page: `https://polyu-vclab.github.io/DepthMaster-page`.

(e.g., Depth Any Panoramas (DAP) [41] for panoramas), they still exhibit noticeable limitations in real-world scenarios. As illustrated by the 3D point cloud reconstructions in Figure 1, these domain-specific specialists often struggle to preserve accurate global geometric structures and fine details. For panoramic specialists, this vulnerability stems primarily from the acute scarcity of panoramic training data, which severely restricts their generalization capabilities. More critically, their specialized, distortion-aware architectures are inherently incompatible with standard perspective inputs, precluding a unified solution to perspective and panorama images.

Consequently, developing a universal metric MDE model faces two intertwined challenges. (1) The first challenge is geometric generalization [48, 20]. A universal model must reconcile the linear projection of pinhole cameras with the non-linear, spherical projection of panoramic cameras without adopting distinct processing pipelines. (2) Another, and perhaps more fundamental, obstacle is data imbalance. While modern perspective MDE methods have flourished by leveraging massive, high-quality datasets, comparable resources for panoramic images are acutely lacking [67, 64]. The key to overcoming this scarcity lies in effectively transferring the rich perspective priors to the panoramic domain. However, existing panoramic methods fail to achieve this because their ad-hoc architectural modifications prevent them from directly inheriting weights and priors from foundation models trained on standard perspective data.

To address these challenges, we introduce **DepthMaster**, a unified monocular metric depth estimation framework. While canonicalizing panoramas into local perspective patches is a common practice to mitigate spherical distortions, prior methods often rely on ad-hoc architectural modifications (e.g., cube convolution [39]) to fuse these patches and handle seam artifacts. Such specialized designs inherently disrupt the standard network structure, hindering the direct transfer of powerful perspective priors learned from large-scale data. Our core innovation lies in unifying panoramic and perspective depth estimation without altering the standard Transformer architecture. Specifically, we project the panoramic image into tangent patches with a 2.5° FoV overlap and inject the spatial layouts of the virtual projection cameras as geometric priors. To ensure global consistency and seamless fusion across patch boundaries, we propose a novel Correspondence Consistency Loss (CCL) as a soft constraint. This loss-based design elegantly substitutes architectural modifications, perfectly preserving the pure backbone. By unifying inputs into this canonical representation and leveraging virtual camera conditioning, DepthMaster accommodates both standard perspective images and panoramic patches, maximizing the utilization of pre-trained perspective priors.

To validate our approach, we perform comprehensive zero-shot metric evaluations across multiple datasets, encompassing both perspective and 360° panoramic domains. The results demonstrate that DepthMaster not only decisively outperforms existing universal methods, such as Unik3D [48] and

DAC [20], but also excels specialist models in their dedicated domains. Specifically, it outperforms leading metric estimators like Depth Anything [85, 86] on perspective imagery and HUSH [32] on panoramic scenes. This dual superiority establishes DepthMaster as the new SOTA in unified monocular metric depth estimation. Our contributions are summarized as follows:

- We propose **DepthMaster**, a unified monocular metric depth estimation framework. By injecting the spatial layouts of virtual projection cameras as geometric priors, we effectively bridge the geometric gap between perspective and panoramic domains without disrupting the standard Transformer architecture.
- We introduce a Correspondence Consistency Loss (CCL) as a soft constraint applied to overlapping patches (e.g., 2.5° FoV overlap). This loss elegantly enforces global consistency and local patch fusion, avoiding ad-hoc architectural modifications and maximizing the inheritance of powerful pre-trained perspective priors.
- DepthMaster achieves **SOTA zero-shot performance** on 13 diverse datasets, outperforming both universal methods and specialist models on their respective benchmarks.

2 Related Work

Monocular Depth Estimation. Early learning-based depth estimation and enhancement methods [13, 15, 93, 84, 83] mainly focus on in-domain settings and lack zero-shot generalization across diverse scenes. To enhance zero-shot generalization, recent works can be partitioned into two main categories. The first scales up model capacity and trains on massive, diverse datasets [51, 88, 52], such as the recent SOTA estimators DepthAnything [85, 86], Metric3D [89, 22], and MoGe [72, 73] trained with millions of samples. However, these methods are designed basically for perspective camera images. The second category re-purposes pre-trained diffusion priors [79, 29, 16] for depth estimation. Our work aligns with the first category, but expands its scope to encompass diverse camera types. By canonicalizing panoramas into overlapping perspective patches and fusing them via a pure loss-based constraint (CCL), we elegantly extend standard perspective foundation models to the panoramic domain without architectural disruption.

Depth Estimation for Panorama. How to handle ERP distortions is the primary challenge for panoramic depth estimation and completion [59, 2, 80, 32, 82, 81]. While early methods adapted convolutional kernels [58, 78] or applied Transformers globally [37, 56, 36], recent approaches project ERP into less distorted representations. However, methods like BiFuse [64] (dual-branch ERP-cubemap fusion), FoVA-Depth [39] (customized CubeConv), and OmniFusion [37] (18 tangent images) rely on specialized, ad-hoc architectural modifications. Such designs inherently disrupt the standard network backbone, hindering joint training with large-scale perspective data and blocking the direct inheritance of powerful pre-trained perspective priors. In contrast, our approach resolves boundary and fusion issues using virtual camera conditioning together with a soft constraint (CCL) without structural modifications, effectively exploiting perspective priors and boosting panoramic performance.

Unified Monocular Depth Estimation. Recent advances seek to reconcile diverse imaging geometries. UniK3D [48] constructs a spherical 3D representation, but its reliance on scarce panoramic data limits generalization. Conversely, DAC [20] establishes ERP as the canonical representation by projecting perspective images into ERP patches. This strategy suffers from two drawbacks: it requires ground-truth camera intrinsics during inference, precluding in-the-wild application, and forcing perspective data into the distorted ERP domain impedes the transfer of perspective priors. In contrast, we unify all inputs within the canonical perspective domain. We decompose panoramas into overlapping perspective patches and inject known virtual spatial layouts, requiring zero ground-truth camera metadata. By preserving the native standard backbone and enforcing global consistency via CCL, our approach overcomes data scarcity and achieves SOTA performance across both benchmarks.

3 Method

We introduce **DepthMaster**, a unified metric depth estimation framework for perspective and 360° cameras. Instead of relying on task-specific operators tailored to cubemap topology, we resolve geometric discrepancies by first canonicalizing any input image into a set of perspective-like patches

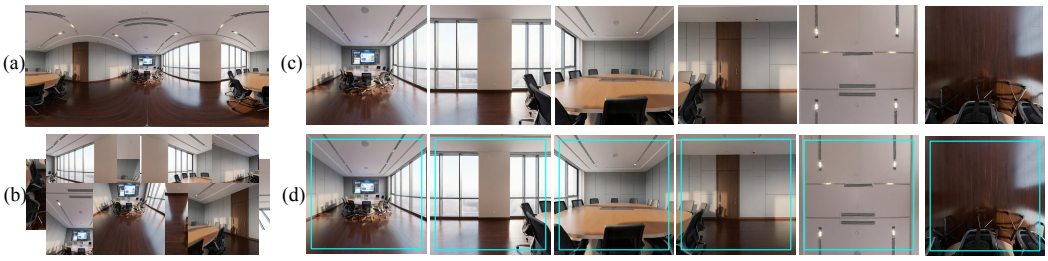


Figure 2: **Comparison of different projection strategies.** (a) Equirectangular. (b) Tangent projection with 18 overlapping patches [37]. (c) Standard 90° Cubemap [64]. (d) Our proposed 95° Cubemap with 2.5° overlap on each side.

(section 3.1), then injecting their virtual spatial layouts as geometric conditions and processing them with a standard Transformer network (section 3.2). Finally, we fuse these patches using our proposed CCL loss (section 3.3), which acts as a soft geometric constraint to preserve backbone purity.

3.1 Canonicalization to Perspective Patches

The primary challenge in unifying depth estimation for perspective and 360° images lies in their distinct geometric projections—linear versus spherical. To bridge this gap, we employ a geometric canonicalization strategy that transforms spherical signals into standard perspective representations, enabling the direct transfer of powerful metric priors within a unified framework.

Rather than learning spherical distortions directly or employing complex multi-tangent schemes (e.g., 18 images [37]), we adopt a simple, standard multi-view formulation. Using tangent plane projection, we unfold the panoramic image into six canonical perspective patches. As illustrated in Figure 2, standard 90° cubemaps suffer from severe boundary discontinuities. To address this, prior methods like BiFuse [64] and FoVA-Depth [39] are forced to introduce customized architectural modifications (e.g., dual-branch fusion or specialized CubeConv operators) to explicitly handle the topology. However, these specific designs largely alter the network structure, preventing direct inference on perspective images and blocking the reuse of pre-trained weights.

To avoid disrupting the backbone while addressing boundary artifacts, we consider expanding the FoV of each patch. Two factors motivate keeping this overlap minimal: first, expanding the FoV well beyond 90° introduces pinhole-violating distortion, which weakens the transfer of perspective priors (especially the metric scale prior); second, an overly large FoV reduces the pixel share of scene objects within each face, degrading fine-grained depth details. Expanding each side by only 2.5° (i.e., a 95° FoV) introduces negligible distortion, as illustrated in Figure 2, while providing sufficient correspondences for cross-patch supervision. The virtual focal length is accordingly adjusted to $f = (w/2)/\tan(95^\circ/2)$ for a patch width w, establishing a shared intrinsic matrix K. This minimal FoV expansion generates the critical overlapping regions required for our CCL loss (section 3.3). By standardizing the inputs into perspective patches with overlaps, this strategy bypasses domain-specific operators, better preserves compatibility with pre-trained perspective priors, and enables a unified inference pipeline across both camera types.

3.2 Framework Architecture

The architecture of DepthMaster is illustrated in Figure 3, which mainly consists of a transformer backbone with camera conditioning, a dense prediction head, and a scale prediction head.

Transformer Backbone. Our architecture is built upon DINOv2 [47] ViT-L/14, a widely adopted backbone for depth estimation tasks that comprises 24 Transformer blocks. Following a design of DA3 [40], the initial 8 layers retain the standard DINOv2 self-attention mechanism. Starting from the 9-th layer, we introduce alternating attention into the remaining 16 blocks. The alternating attention is achieved entirely by reshaping the input tensor dimensions. The frame-wise attention captures intra-patch structure, whereas global attention establishes interactions across patches, thereby facilitating spatial correspondence and geometric coherence. Our canonicalization process generates N perspective patches: a single patch (N=1) for standard perspective images or six patches (N=6) for panoramic inputs. In the single-patch scenario, global attention naturally reduces to standard self-attention, ensuring unified training for both inputs.

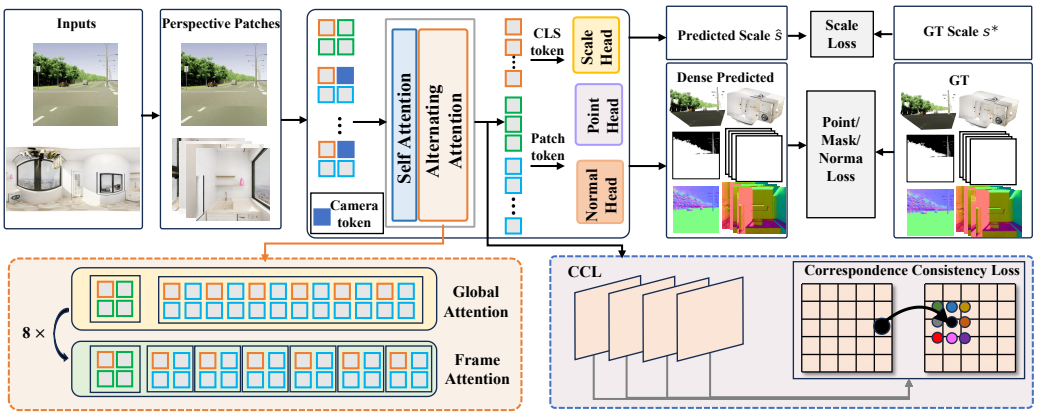


Figure 3: **The pipeline of DepthMaster.** Perspective and panoramic images are unified as perspective patches and processed by a shared transformer with virtual camera conditioning, while CCL explicitly enforces consistency across overlapping patches during training. The CLS token for perspective inputs, or pooled CLS tokens for panoramic inputs, is used to predict a shared global scale.

Virtual Camera Conditioning. For panoramic inputs, we leverage the virtual camera extrinsics and intrinsics used in cubemap projection (section 3.1) as conditioning signals to provide geometric priors.

The world-to-camera extrinsic matrix $\mathbf{T} \in SE(3)$ is first converted to a camera-to-world representation and parameterized as a 7-dimensional vector $\mathbf{p} = [\mathbf{t}; \mathbf{q}] \in \mathbb{R}^7$, where $\mathbf{t} \in \mathbb{R}^3$ is the translation and $\mathbf{q} \in \mathbb{R}^4$ is the unit quaternion encoding rotation. This vector is embedded via a two-layer MLP with SiLU activation:

$$\mathbf{e}_{\text{pose}} = \text{MLP}_{\text{pose}}(\mathbf{p}) \in \mathbb{R}^d, \tag{1}$$

and prepended to the token sequence as a special conditioning token.

The camera intrinsics are summarized as a 4-dimensional vector $\mathbf{r} = [f_x/W, f_y/H, c_x/W, c_y/H]$, encoding the normalized focal lengths and principal point, which is embedded via a two-layer MLP:

$$\mathbf{e}_{\text{ray}} = \text{MLP}_{\text{ray}}(\mathbf{r}) \in \mathbb{R}^d, \tag{2}$$

and prepended as another special token. For perspective image inputs, no camera conditioning is applied, allowing the model to jointly train on perspective images without requiring camera metadata.

Dense Prediction Heads. We have two dense prediction heads sharing the same Dense Prediction Transformer (DPT) architecture [51], which takes multi-scale features from layers $\{11, 15, 19, 23\}$ of the backbone. The point map head outputs a 4-channel dense prediction $\hat{\mathbf{P}} \in \mathbb{R}^{H\times W\times 4}$. The first three channels represent the 3D point map in world coordinates for panoramic inputs or affine camera space for perspective inputs. The fourth channel predicts a single-channel validity mask $\hat{\mathbf{M}} \in [0,1]^{H\times W}$ indicating non-sky regions. For the 3D point map, we supervise it using affine-invariant point cloud loss. For each sphere scale α, we sample anchor points $\mathcal{H}_\alpha$ and compute:

$$\mathcal{L}_{S(\alpha)} = \sum_{j\in\mathcal{H}_\alpha} \sum_{i\in S_j} \frac{1}{z_i} \left\| s_j^* \hat{\mathbf{p}}_i + \mathbf{t}_j^* - \mathbf{p}_i \right\|_1, \tag{3}$$

where we apply the ROE alignment solver [72] to obtain $(s_j^*, \mathbf{t}_j^*)$ with $\mathbf{t}_j^* = (0, 0, t_{z,j}^*)^\top$. Specifically, $t_{x,j}^* = t_{y,j}^* = 0$ and only an axial offset $t_{z,j}^*$ is permitted, ensuring that the predicted point map remains constrained along camera rays without lateral drift. Because perspective training images are provided without camera intrinsics and exhibit inherent focal-distance (dolly-zoom) ambiguity, this unified affine-invariant objective absorbs local axial depth bias and allows the shared point-map head to focus on relative 3D structure across both perspective and panoramic branches, while the absolute metric scale is learned separately by the scale prediction head. We use three scales: $\alpha \in \{1/4, 1/16, 1/64\}$, denoted as $\mathcal{L}_{S_1}$, $\mathcal{L}_{S_2}$, and $\mathcal{L}_{S_3}$. For the mask channel, we supervise it with the mask BCE loss:

$$\mathcal{L}_M = \|\hat{\mathbf{M}} - (1 - \mathbf{M}_{\text{inf}})\|_2^2, \tag{4}$$

where $\mathbf{M}_{\text{inf}}$ is the infinity mask label obtained from ground truth for synthetic data or SegFormer [77] for outdoor scenes. The surface normal head outputs an ℓ_2-normalized 3-channel normal map $\hat{\mathbf{N}} \in \mathbb{R}^{H\times W\times 3}$ with unit vectors $\hat{\mathbf{n}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2}$. We supervise it with the following normal loss:

$$\mathcal{L}_N = \sum_{i\in\mathcal{M}} \angle(\hat{\mathbf{n}}_i, \mathbf{n}_i), \tag{5}$$

where $\hat{\mathbf{n}}_i$ is the predicted unit normal from the normal head, and $\angle(\cdot,\cdot)$ measures angular difference.

Scale Prediction Head. In metric depth prediction approaches such as Metric3D [22] and MoGe v2 [73], scale is typically estimated separately. To avoid compromising the accuracy of relative geometry, we introduce a dedicated branch with independent supervision for scale prediction:

$$\mathcal{L}_s = \|\log(\hat{s}) - \text{stopgrad}(\log(s^*))\|_2^2, \tag{6}$$

where $\log(\hat{s})$ denotes the predicted metric scale in log space, and s^* is the optimal scale computed on-the-fly between the predicted scale-invariant depth map $\hat{d}$ and the ground truth. To ensure strict scale consistency across multi-patch panoramic inputs, we perform average pooling over the CLS tokens of all patches to form a unified global representation before passing it to the scale predictor. By predicting a single, globally shared scale factor and multiplying it with the scale-invariant point maps, our framework inherently avoids scale discontinuities or drift at the patch boundaries.

3.3 Correspondence Consistency Loss

Although global attention facilitates information exchange across patches, it does not explicitly enforce consistency within overlapping regions. To address this limitation, we introduce a novel loss function, namely Correspondence Consistency Loss (CCL), grounded in dense patch correspondences.

Building Correspondences. For each pair of overlapping patches (i, j), we construct a dense correspondence field based on their relative orientations. Each patch is treated as a virtual perspective camera sharing the same intrinsic matrix K but having its own rotation matrix R_k, which maps points from the camera coordinate system to the world coordinate system. Since we only consider directions at infinity, the mapping between patches i and j can be determined by:

$$\mathbf{H}_{ij} = K\, R_j^{-1}\, R_i\, K^{-1}, \tag{7}$$

which warps a pixel $\mathbf{p}_i$ in patch i to its corresponding location $\mathbf{p}_j \sim \mathbf{H}_{ij}\mathbf{p}_i$ in patch j. Because $\mathbf{p}_j$ often lies at non-integer coordinates, directly interpolating the final predicted 3D point cloud near depth discontinuities can mix foreground and background points, resulting in large geometric errors or "flying points." Instead, applying CCL at the feature level provides a softer, high-dimensional structural consistency constraint while avoiding directly penalizing 3D artifacts caused by point-cloud interpolation. Thus, the feature representation $\mathbf{F}_j^m(\mathbf{p}_j)$ at backbone layer m is obtained via differentiable bilinear interpolation from the feature map $\mathbf{F}_j^m$.

Loss Formulation. The CCL is applied to multi-scale features from layers $\{11, 15, 19, 23\}$ of the backbone. It aggregates the L_2 distance between the corresponding feature vectors over all valid correspondences, overlapping pairs $\mathcal{O}$, and selected backbone layers:

$$\mathcal{L}_{\text{CCL}} = \sum_{m=1}^{M} \sum_{(i,j)\in\mathcal{O}} \frac{1}{|\Omega_{ij}|} \sum_{(\mathbf{p}_i,\mathbf{p}_j)\in\Omega_{ij}} \left\|\mathbf{F}_i^m(\mathbf{p}_i) - \mathbf{F}_j^m(\mathbf{p}_j)\right\|_2^2, \tag{8}$$

where M is the number of backbone layers to which the loss is applied, and Ω_{ij} denotes the set of valid corresponding pixel pairs $(\mathbf{p}_i, \mathbf{p}_j)$ in the 2.5° overlapping region between adjacent patches i and j, as determined by the geometric projection in eq. (7). Rather than imposing spatial smoothing over neighboring pixels, CCL only aligns geometrically corresponding locations across overlapping patches that observe the same spherical direction, thereby encouraging consistent representations without boundary over-smoothing. Moreover, since CCL supervises high-level features that participate in global attention, this consistency constraint propagates beyond the local overlap regions and improves global geometric coherence across the panorama.

4 Experiments

4.1 Experimental Settings

Experiments Setup. Our model is trained on 15 datasets, including 14 perspective datasets and 1 panoramic (i.e., Structured3D [96]). For real-world data, due to inherent noise (e.g., incompleteness, blurred object boundaries, and severe flying pixels), we refine the ground-truth depth maps. To evaluate the zero-shot generalization capabilities of our DepthMaster model, we test it on 13 unseen datasets (10 perspective and 3 panoramic). It is worth mentioning that **our model is trained on a significantly smaller data scale compared to existing models** such as MoGe [72], DepthAnything [85], and the unified model UniK3D [48] (see Appendix C.4). Detailed dataset statistics, implementation details, and real data refinement strategies are provided in Appendix C and A.1, respectively.

Evaluation Details. We conduct comprehensive evaluations of DepthMaster's zero-shot generalization capabilities across both perspective and panoramic images using a **single pre-trained model without any domain-specific fine-tuning**. For perspective images, we follow the MoGe protocol [72]. For panoramic images, evaluations are performed on the full ERP maps following the DA^2 [36] protocol. While our method directly outputs a unified 3D point cloud in the world coordinate system, to ensure a fair comparison with existing baselines, we reproject the predicted 95° FoV cubemap back into a 2D ERP depth map format, applying a simple weighted average to merge the overlapping regions. Comprehensive evaluation protocols and metrics are detailed in Appendix A.3.

4.2 Main Results

Results on Panoramic Images. We evaluate DepthMaster on three panoramic benchmarks: Stanford2D3D [3], Matterport3D [9], and PanoSUNCG [63]. As shown in Table 1, DepthMaster achieves remarkable zero-shot performance across all three evaluation protocols, surpassing both specialized and unified models. For scale-invariant depth, DepthMaster substantially outperforms the recent zero-shot model DA^2 [36], reducing the average AbsRel from 6.05 to 4.77 and boosting δ_1 accuracy from 95.63% to 96.19%, while also surpassing in-domain panoramic specialists such as Elite360D [2] and the multi-tangent fusion method OmniFusion [37] without any iterative post-refinement. For affine-invariant depth, DepthMaster further outperforms the SOTA specialized model DA360 [27], cutting the average AbsRel from 6.39 to 4.58 and improving δ_1 from 94.53% to 96.33%. Critically, DepthMaster's zero-shot performance on Matterport3D surpasses UniK3D (5.79 vs. 8.14 AbsRel) by a large margin, although UniK3D is explicitly trained on this dataset. For metric depth, DepthMaster also consistently outperforms the SOTA metric depth model DAP [41] (7.25 vs. 11.79 AbsRel). It is worth mentioning that DepthMaster achieves these results by **training only on a single panoramic dataset**, demonstrating its exceptional data efficiency.

Results on Perspective Images. We further evaluate DepthMaster's performance on ten diverse perspective datasets: NYUv2 [57], KITTI [61], ETH3D [54], iBims-1 [31, 30], GSO [12], Sintel [7], DDAD [18], DIODE [62], Spring [45], and HAMMER [28]. We benchmark our method against strong specialist baselines across both relative and metric geometry settings. As shown in Table 2, **DepthMaster achieves the best overall average rank (1.31), outperforming the previous strongest method, MoGe 2 (2.22), by a large margin**. Specifically, our method delivers the best average results in scale-invariant point prediction, affine-invariant point prediction, metric depth prediction, scale-invariant depth prediction, affine-invariant depth prediction, and affine-invariant disparity prediction. For metric point prediction, it achieves the top δ_1 score (93.9%), while remaining competitive in AbsRel (8.27). These results demonstrate DepthMaster's strong generalization across diverse perspective scenes, despite using only one-fourth the training data of recent unified models (see Appendix C.4).

Qualitative Comparison. As shown in Figure 4 and Figure 5, DepthMaster generates depth maps with superior fidelity and structural consistency. It effectively recovers fine details and preserves sharp boundaries, whereas competing unified models like UniK3D often produce over-smoothed details and visible artifacts. These qualitative results align with our quantitative findings, further validating the effectiveness of our design. Due to space constraints, we provide additional qualitative comparisons in Appendix D to demonstrate the strong generalization ability. We strongly encourage readers to explore the interactive demos on our project page: `https://polyu-vclab.github.io/DepthMaster-page`.

Table 1: Quantitative results on three panorama datasets. The best and second best results are highlighted in red and blue, respectively. Gray numbers denote models trained on this dataset. Unavailable/unsupported metrics are marked with '-'. Methods marked with † are in-domain models trained on the corresponding panorama datasets, and are listed for reference only (not ranked). Results in the Scale-invariant depth map section are taken from DA^2 [36]. Our model surpasses not only other unified models but also all specialized models designed specifically for panoramic imagery.

Method	Stanford2D3D				Matterport3D				PanoSUNCG				*Avg.*	
	AbsRel↓	RMSE↓	δ_1 ↑	δ_2 ↑	AbsRel↓	RMSE↓	δ_1 ↑	δ_2 ↑	AbsRel↓	RMSE↓	δ_1 ↑	δ_2 ↑	AbsRel↓	δ_1 ↑
Scale-invariant depth map														
OmniDepth† [98]	19.96	61.52	68.77	88.91	29.01	76.43	68.30	87.94	11.43	37.10	87.05	93.65	20.13	74.71
BiFuse† [64]	12.09	41.42	86.60	95.80	20.48	62.59	84.52	93.19	5.92	25.96	95.90	98.23	12.83	89.01
BiFuse++† [65]	11.17	37.20	87.83	96.49	14.24	51.90	87.90	95.17	5.24	24.77	96.30	98.35	10.22	90.68
UniFuse† [26]	11.14	36.91	87.11	96.64	10.63	49.41	88.97	96.23	5.28	27.04	95.91	98.25	9.02	90.66
Elite360D† [2]	11.82	37.56	88.72	96.84	11.15	48.75	88.15	96.46	–	–	–	–	–	–
OmniFusion† [37]	9.50	34.74	89.88	97.69	9.00	42.61	91.89	97.97	–	–	–	–	–	–
PanoFormer† [56]	11.31	35.57	88.08	96.23	9.04	44.70	88.16	96.61	5.34	18.90	94.87	98.83	8.56	90.37
HRDFuse† [1]	9.35	31.06	91.40	97.98	9.67	44.33	91.62	96.69	6.90	27.44	92.15	97.42	8.64	91.72
DepthAnywhere† [67]	11.80	35.10	91.00	97.10	8.50	–	91.70	97.60	–	–	–	–	–	–
PanDA [8]	48.44	53.06	33.92	51.33	37.10	101.5	42.51	67.29	34.73	79.69	44.49	71.45	40.09	40.31
UniK3D [48]	11.31	19.72	88.94	95.33	9.66	32.66	93.00	98.58	11.46	25.38	90.18	98.02	11.39	89.56
DA^2 [36]	6.09	19.09	95.94	98.51	6.83	27.51	94.89	97.71	5.22	19.31	96.06	98.28	6.05	95.63
DepthMaster (Ours)	4.81	17.63	96.15	98.56	5.95	24.71	94.92	98.13	3.54	15.46	97.50	98.81	4.77	96.19
Affine-invariant depth map														
DAP [41]	6.80	22.69	93.92	98.04	8.98	35.03	89.68	96.16	5.53	24.55	95.54	97.88	7.10	93.05
UniK3D [48]	9.02	22.72	92.92	98.45	8.14	28.80	94.34	97.87	8.06	19.25	94.02	99.13	8.41	93.76
DA360 [27]	6.55	21.13	94.93	97.59	7.92	32.75	91.39	96.00	4.70	20.29	97.26	98.24	6.39	94.53
DepthMaster (Ours)	4.59	17.59	96.27	98.56	5.79	24.68	94.99	98.14	3.36	15.32	97.74	98.99	4.58	96.33
Metric depth map														
DAP [41]	10.89	28.45	91.44	98.03	17.78	83.40	81.92	93.78	6.71	25.12	94.84	97.69	11.79	89.40
UniK3D [48]	22.14	45.75	68.79	97.05	21.74	54.51	69.74	96.32	21.69	39.89	71.48	95.45	21.86	70.00
DepthMaster (Ours)	7.73	20.04	95.58	98.55	9.49	29.85	91.65	97.65	4.54	15.99	97.34	98.76	7.25	94.86

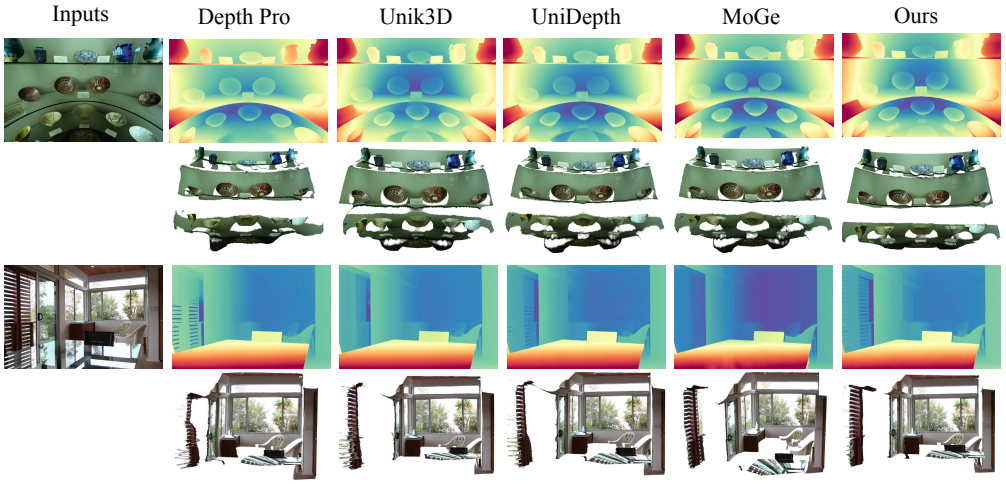


Figure 4: **Qualitative comparison on perspective images.** Compared with state-of-the-art models (Depth Pro, UniK3D, UniDepth, and MoGe), DepthMaster demonstrates a robust capability to recover fine structures, preserve sharp boundaries, and maintain accurate geometric consistency.

Impact of CCL Loss. Removing $\mathcal{L}_{\text{CCL}}$ has a negligible effect on perspective images but raises the panoramic AbsRel from 6.31 to 7.41 (δ_1: 93.3%,→,92.4%). As illustrated in Figure 6, conspicuous geometric inconsistencies and visible seams appear in the overlapping regions. This confirms that CCL is essential for merging patch-level estimates into a globally coherent depth map on panorama inputs.

Impact of Virtual Camera Pose. Omitting camera pose conditioning only mildly affects perspective performance (AbsRel 7.15,→,7.47), but substantially degrades panoramic results (AbsRel

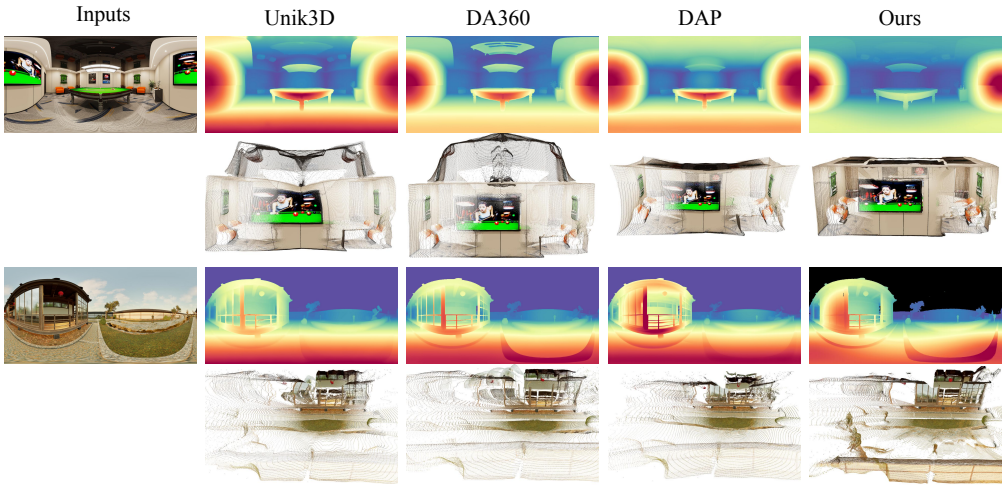


Figure 5: **Qualitative comparison on panoramic images.** Compared with UniK3D, DA360, and DAP, our method excels at maintaining geometric accuracy and is free of the severe distortion artifacts commonly encountered in other methods.

Table 2: Quantitative evaluation for *metric and relative geometry* across 10 zero-shot perspective benchmarks (NYUv2, KITTI, ETH3D, iBims-1, GSO, Sintel, DDAD, DIODE, Spring, and HAMMER), comparing specialized and unified models. The best and second best results are highlighted in red and blue, respectively. Gray numbers denote models trained on this dataset. We adopt the results reported in MoGe [72] and MoGe V2 [73] for all methods, excluding the unified model. Unavailable/unsupported metrics are marked with '–'. The complete quantitative results for each individual dataset are provided in Appendix B.

Method	Point								Depth								*Avg.*
	Metric		Scale-inv.		Affine-inv.		Local		Metric		Scale-inv.		Affine-inv.		Affine-inv. (disp)		
	$\text{Rel}^{\text{p}}\downarrow$	$\delta_1^{\text{p}}\uparrow$	$\text{Rel}^{\text{p}}\downarrow$	$\delta_1^{\text{p}}\uparrow$	$\text{Rel}^{\text{p}}\downarrow$	$\delta_1^{\text{p}}\uparrow$	$\text{Rel}^{\text{p}}\downarrow$	$\delta_1^{\text{p}}\uparrow$	$\text{Rel}^{\text{d}}\downarrow$	$\delta_1^{\text{d}}\uparrow$	$\text{Rel}^{\text{d}}\downarrow$	$\delta_1^{\text{d}}\uparrow$	$\text{Rel}^{\text{d}}\downarrow$	$\delta_1^{\text{d}}\uparrow$	$\text{Rel}^{\text{d}}\downarrow$	$\delta_1^{\text{d}}\uparrow$	Rk.↓
ZoeDepth [5]	–	–	–	–	–	–	–	–	39.3	49.9	12.7	83.9	10.1	88.5	11.1	88.3	10.25
DA V1 [85]	–	–	–	–	–	–	–	–	31.8	54.8	11.7	85.8	8.76	90.4	8.63	92.2	8.00
DA V2 [86]	–	–	–	–	–	–	–	–	29.9	56.6	10.7	87.6	8.48	90.8	8.82	91.6	7.13
Metric3D V2 [22]	–	–	–	–	–	–	–	–	–	–	7.92	91.8	7.66	92.9	9.51	89.4	5.67
MASt3R [34]	26.2	55.3	14.5	82.1	11.6	86.0	8.09	92.2	49.7	30.3	11.2	86.5	9.38	89.1	11.6	87.8	8.25
UniDepth V1 [50]	12.1	87.2	13.6	85.0	10.9	88.1	9.21	91.0	23.2	67.5	10.1	89.1	8.61	91.0	9.75	89.9	6.44
Depth Pro [6]	13.7	81.9	12.4	87.7	9.93	89.4	6.91	94.1	27.6	54.4	9.81	89.1	7.65	92.0	8.42	91.7	5.50
UniDepth V2 [49]	10.1	91.9	11.6	87.7	8.56	91.9	6.34	94.9	21.3	75.3	8.61	90.8	6.42	93.9	7.35	93.0	3.75
UniK3D [48]	8.18	92.3	10.7	88.9	8.18	92.3	6.25	94.9	15.5	82.5	8.26	91.1	6.50	93.8	7.38	93.0	2.75
MoGe 2 [73]	8.19	93.6	10.8	88.5	7.98	91.7	5.33	95.9	15.7	76.8	7.35	92.2	5.62	94.8	6.66	93.8	2.22
Ours	8.27	93.9	8.12	92.1	6.18	93.9	5.34	96.0	14.2	82.3	6.76	92.7	5.18	94.9	6.80	93.8	1.31

Table 3: Ablation study on the key components in DepthMaster, reporting affine-invariant metrics averaged over all 10 perspective and 3 panoramic benchmarks.

Model Variant	Perspective		Panoramic	
	Rel↓	$\delta_1\uparrow$	Rel↓	$\delta_1\uparrow$
Naive Baseline	7.74	91.0	9.45	89.7
w/o CCL	7.28	91.7	7.41	92.4
w/o Camera Pose	7.47	91.4	14.36	85.8
Panoramic-Only	16.34	78.1	8.28	91.5
DepthMaster (Full)	**7.15**	**91.8**	**6.31**	**93.3**

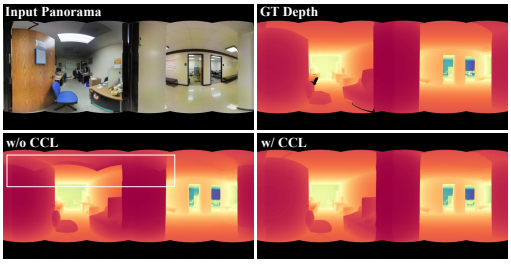


Figure 6: Impact of the proposed CCL loss on depth prediction quality.

6.31,→,14.36, δ_1 93.3%,→,85.8%). Explicit pose conditioning is therefore crucial for resolving the complex spatial geometry of panoramic scenes.

Impact of Unified Perspective Representation. We train a "Naive" baseline that removes the canonicalization module and directly feeds a mixture of perspective and ERP inputs to the network.

Table 4: Efficiency and accuracy comparison on a single A100 GPU under FP16. Accuracy is the panoramic affine-invariant average across Stanford2D3D, Matterport3D, and PanoSUNCG at 1024×2048 ERP output resolution.

Method	Input	Output Res.	AbsRel↓	$\delta_1 \uparrow$	Runtime (ms)↓	Peak Mem. (GB)↓
UniK3D [48]	ERP (1024×2048)	1024×2048	8.41	93.76	279	5.15
DA360 [27]	ERP (1024×2048)	1024×2048	6.39	94.53	269	6.65
DAP [41]	ERP (1024×2048)	1024×2048	7.10	93.05	218	6.50
DepthMaster (Ours)	Cube (6×518^2)	1024×2048	**4.58**	**96.33**	**178**	**4.58**

Perspective performance degrades moderately (AbsRel 7.15,→,7.74), while panoramic performance drops sharply (AbsRel 6.31,→,9.45, δ_1 93.3%,→,89.7%), confirming that a single network struggles to reconcile fundamentally different projections.

Impact of Training Data. We train a "Panoramic-Only" variant on the single panoramic dataset while keeping the same canonicalization pipeline, so the network still sees cubemap patches. Due to limited scene diversity and the domain gap between panorama-derived patches and natural photographs, perspective performance drops severely (AbsRel 7.15,→,16.34, δ_1 91.8%,→,78.1%), and panoramic performance also degrades (AbsRel 6.31,→,8.28, δ_1 93.3%,→,91.5%). This shows that large-scale perspective data not only preserve perspective generalization but also enhance panoramic depth estimation via cross-domain prior transfer.

Prior Transferability and Efficiency. Our method surpasses the SOTA DAP method [41] with significantly less panoramic data, utilizing only 35K panoramic training samples compared to 1.9M required by DAP. This demonstrates the strong prior transferability of our framework, effectively unlocking perspective priors for the data-scarce panoramic domain. To further evaluate computational efficiency, we benchmark all methods on a single A100 GPU under FP16 with FlashAttention enabled, where all methods output a 1024×2048 ERP depth map (including the CUDA-based cubemap-to-ERP re-projection step for DepthMaster). As shown in Table 4, DepthMaster is not only substantially more accurate, but also achieves faster inference (178 ms) and lower peak GPU memory consumption (4.58 GB) than competing panoramic and unified baselines.

5 Conclusion

In this paper, we introduced DepthMaster, a unified monocular depth estimation framework for both perspective and panoramic images. Our core strategy was to canonicalize any input image into standard perspective patches, thereby overcoming the challenges of geometric variance and data scarcity of panoramic cameras. A novel correspondence consistency loss was proposed to ensure geometric consistency across these patches. Extensive evaluations demonstrated that DepthMaster achieved new state-of-the-art performance on 13 diverse datasets of different camera types, surpassing existing universal and specialized methods. Our work presented a simple, powerful, and scalable paradigm for building universal geometric perception systems.

Limitations. While effective, DepthMaster still has some limitations. First, our canonicalization and virtual camera conditioning are specifically designed for perspective and 360° panoramic images. For other heavily distorted cameras such as fisheye lenses, images with known intrinsics can be mapped onto the unit sphere and resampled into cubemap faces, whereas uncalibrated fisheye images with unknown intrinsics cannot be directly processed without prior calibration. Second, generalization to highly abstract out-of-distribution (OOD) data remains a common challenge to DepthMaster and almost all state-of-the-art models. Like existing methods, DepthMaster encounters difficulties with artistic images that deviate significantly from the training distribution, such as ink wash paintings (see Appendix E for examples). In the future, we plan to extend our unified framework to broader uncalibrated camera models and leverage style-transfer data synthesis to improve robustness on abstract domains.

Acknowledgments and Disclosure of Funding

This work is supported by the PolyU-OPPO Joint Innovative Research Center.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13273–13282, 2023.

[2] Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *CVPR*, 2024.

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

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[6] 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.

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012.

[8] Zidong Cao, Jinjing Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, and Lin Wang. Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 982–992, 2025.

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676. IEEE Computer Society, 2017.

[10] Liyi Chen, Ruihuang Li, Guowen Zhang, Pengfei Wang, and Lei Zhang. Fast multi-view consistent 3d editing with video priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 2948–2956, 2026.

[11] Liyi Chen, Pengfei Wang, Guowen Zhang, Zhiyuan Ma, and Lei Zhang. Omni-3dedit: Generalized versatile 3d editing in one-pass. *arXiv preprint arXiv:2603.17841*, 2026.

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items, 2022.

[13] 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.

[14] Gasser Elazab, Frank Neuhaus, Tilman Koß, Malte Splietker, Aditya Date, Michael Unterreiner, Maximilian Jansen, and Olaf Hellwich. Card: A multi-modal automotive dataset for dense 3d reconstruction in challenging road topography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17820–17830, 2026.

[15] 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.

[16] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024.

[17] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. *arXiv preprint arXiv:2004.06320*, 2020.

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

[19] Yanjun Guo, Zhengqiang Zhang, Pengfei Wang, Xinyue Liang, Zhiyuan Ma, and Lei Zhang. Memorize when needed: Decoupled memory control for spatially consistent long-horizon video generation. *arXiv preprint arXiv:2604.18215*, 2026.

[20] Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26996–27006, 2025.

[21] 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, 2023.

[22] 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. *arXiv preprint arXiv:2404.15506*, 2024.

[23] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[24] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024.

[25] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021.

[27] Hualie Jiang, Ziyang Song, Zhiqiang Lou, Rui Xu, and Minglang Tan. Depth anything in 360°: Towards scale invariance in the wild. *arXiv preprint arXiv:2512.22819*, 2025.

[28] HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, et al. On the importance of accurate geometry data for dense 3d vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–791, 2023.

[29] 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.

[30] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Körner. Evaluation of cnn-based single-image depth estimation methods. In Laura Leal-Taixé and Stefan Roth, editors, *Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS)*, pages 331–348. Springer International Publishing, 2019.

[31] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020.

[32] Jongsung Lee, Harin Park, Byeong-Uk Lee, and Kyungdon Joo. Hush: Holistic panoramic 3d scene understanding using spherical harmonics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16599–16608, 2025.

[33] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024.

[34] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.

[35] Boyi Li, Yifan Shen, Yuanzhe Liu, Yifan Xu, Jiateng Liu, Xinzhuo Li, Zhengyuan Li, Jingyuan Zhu, Yunhan Zhong, Fangzhou Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[36] Haodong Li, Wangguangdong Zheng, Jing He, Yuhao Liu, Xin Lin, Xin Yang, Ying-Cong Chen, and Chunchao Guo. Da^2: Depth anything in any direction. *arXiv preprint arXiv:2509.26618*, 2025.

[37] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022.

[38] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Daniel Lichy, Hang Su, Abhishek Badki, Jan Kautz, and Orazio Gallo. FoVA-Depth: Field-of-view agnostic depth estimation for cross-dataset generalization. In *International Conference on 3D Vision (3DV)*, 2024.

[40] Haotong Lin, Sili Chen, Junhao 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.

[41] Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. *arXiv*, 2025.

[42] Yuanzhe Liu, Jingyuan Zhu, Yuchen Mo, Gen Li, Xu Cao, Jin Jin, Yifan Shen, Zhengyuan Li, Tianjiao Yu, Wenzhen Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024.

[44] Zhiyuan Ma, Wenbo Hu, Wang Zhao, Pengfei Wang, Ying Shan, and Lei Zhang. Oreo: Fidelity alignment in 3d generation via on-the-fly rendering-editing optimization. In *European Conference on Computer Vision*, pages 527–544. Springer, 2026.

[45] 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 *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[46] Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 3d ken burns effect from a single image. *ACM Transactions on Graphics*, 38(6):184:1–184:15, 2019.

[47] 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.

[48] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

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

[54] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[55] Yifan Shen, Yuanzhe Liu, Jingyuan Zhu, Xu Cao, Xiaofeng Zhang, Yixiao He, Wenming Ye, James Rehg, and Ismini Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

[56] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360° depth estimation. In *European Conference on Computer Vision*, pages 195–211. Springer, 2022.

[57] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[58] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[59] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2573–2582, 2021.

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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

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

[63] Fu-En Wang, Hou-Ning Hu, Hsien-Tzu Cheng, Juan-Ting Lin, Shang-Ta Yang, Meng-Li Shih, Hung-Kuo Chu, and Min Sun. Self-supervised learning of depth and camera motion from 360 videos. In *Asian Conference on Computer Vision*, pages 53–68. Springer, 2018.

[64] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 462–471, 2020.

[65] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5448–5460, 2022.

[66] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *CoRR*, abs/1909.05452, 2019.

[67] Ning-Hsu Albert Wang and Yu-Lun Liu. Depth anywhere: Enhancing 360 monocular depth estimation via perspective distillation and unlabeled data augmentation. *Advances in Neural Information Processing Systems*, 37:127739–127764, 2024.

[68] Pengfei Wang, Liyi Chen, Zhiyuan Ma, Yanjun Guo, Guowen Zhang, and Lei Zhang. One2scene: Geometric consistent explorable 3d scene generation from a single image. *arXiv preprint arXiv:2602.19766*, 2026.

[69] Pengfei Wang, Yuxi Wang, Shuai Li, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Open vocabulary 3d scene understanding via geometry guided self-distillation. In *European Conference on Computer Vision*, pages 442–460. Springer, 2024.

[70] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. IRS: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *CoRR*, abs/1912.09678, 2019.

[71] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.

[72] 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.

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

[74] 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)*, 2020.

[75] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

[76] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[77] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[78] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 2024.

[79] Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv preprint arXiv:2403.06090*, 2024.

[80] Qingsong Yan, Qiang Wang, Kaiyong Zhao, Jie Chen, Bo Li, Xiaowen Chu, and Fei Deng. Spherefusion: Efficient panorama depth estimation via gated fusion. In *International Conference on 3D Vision 2025*, 2025.

[81] Zhiqiang Yan, Xiang Li, Kun Wang, Shuo Chen, Jun Li, and Jian Yang. Distortion and uncertainty aware loss for panoramic depth completion. In *International Conference on Machine Learning*, pages 39099–39109. PMLR, 2023.

[82] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022.

[83] Zhiqiang Yan, Zhengxue Wang, Haoye Dong, Jun Li, Jian Yang, and Gim Hee Lee. Ducos: Duality constrained depth super-resolution via foundation model. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8361–8371. IEEE, 2025.

[84] Zhiqiang Yan, Zhengxue Wang, Kun Wang, Jun Li, and Jian Yang. Completion as enhancement: A degradation-aware selective image guided network for depth completion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26943–26953. IEEE, 2025.

[85] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024.

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

[87] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[88] Wei Yin, Yifan Liu, and Chunhua Shen. Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7282–7295, 2021.

[89] 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.

[90] 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. *CoRR*, abs/2012.09365, 2020.

[91] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Yuanzhe Liu, and Ismini Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025.

[92] Tianjiao Yu, Xinzhuo Li, Yifan Shen, Onkar Susladkar, Yuanzhe Liu, Xiaona Zhou, and Ismini Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[93] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3916–3925, 2022.

[94] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469, 2020.

[95] Amir R Zamir, Alexander Sax, William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[96] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020.

[97] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19855–19865, 2023.

[98] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018.

Appendix

In this appendix, we provide the following materials:

- **A** Implementation details, covering training configurations and the unified evaluation protocols adopted for both perspective and panoramic datasets (referring to Sec. A).
- **B** Complete the evaluation results on individual perspective benchmarks, offering a granular view beyond the aggregated metrics reported in the main paper (referring to Sec. B).
- **C** Detailed descriptions of the training and validation datasets used in our experiments, including their scale, modality, and sampling ratios (referring to Sec. C).
- **D** More qualitative comparisons with state-of-the-art methods across diverse indoor, outdoor, and panoramic scenes (referring to Sec. D).
- **E** Discussion of current limitations of DepthMaster and potential future research directions (referring to Sec. E).
- **F** Discussion of the broader impacts (referring to Sec. F).
- **G** Asset licenses and terms of use for all datasets and pre-trained models employed in this work (referring to Sec. G).

A Implementation Details

A.1 Training Details

We base our model architecture on DINOv2 [47] and initialize with its pre-trained weights. For optimization, we employ the AdamW optimizer with a cosine decay learning rate schedule, setting the learning rate to 5×10^{-5}. During the training phase, we uniformly resize the input images to a width of 518 pixels, followed by cropping to achieve an aspect ratio uniformly distributed in the range of [1.0, 2.0]. The model is trained for 200K iterations on 32 NVIDIA H100 GPUs with a total batch size of 256. To enhance the model's robustness, we employ a series of data augmentation techniques, including color jittering, Gaussian blur, and JPEG compression artifacts. Given that our training data includes only one panorama dataset, we assign it a fixed sampling ratio of 20% to ensure its adequate representation during training.

Loss Weights. The total training objective is a weighted combination of the point map loss, the surface normal loss, the metric scale loss, and the validity mask loss. Specifically, we set the weights of the point map loss and the surface normal loss to 1.0, and the weights of the metric scale loss and the validity mask loss to 0.1, i.e.,

$$\mathcal{L} = 1.0 \cdot \mathcal{L}_{\text{point}} + 1.0 \cdot \mathcal{L}_{\text{normal}} + 0.1 \cdot \mathcal{L}_{\text{scale}} + 0.1 \cdot \mathcal{L}_{\text{mask}}. \tag{A.1}$$

This configuration balances the contributions from the dense geometric supervision (point map and surface normal) and the auxiliary metric and validity supervision, and is kept fixed throughout all training stages.

A.2 Evaluation Protocol

Relative Geometry. To ensure a rigorous and fair evaluation of relative geometry, we adopt the alignment strategy introduced in MoGe [72]. Before calculating the errors, the predictions are aligned with the ground truth data using optimal scale and shift (where applicable) for each individual sample. The specific alignment formulations are detailed as follows:

- *Scale-invariant point map.* We align the predicted point map $\hat{\mathbf{p}}_i$ with the ground truth $\mathbf{p}_i$ by optimizing a global scale factor a^*. The objective is to minimize the depth-weighted L_1 distance:

$$a^* = \underset{a}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \| a\hat{\mathbf{p}}_i - \mathbf{p}_i \|_1. \tag{A.2}$$

- *Affine-invariant point map.* When both scale and translation are involved, we compute the optimal scale a^* and shift $\mathbf{b}^*$ to minimize the alignment error:

$$(a^*, \mathbf{b}^*) = \underset{a,\mathbf{b}}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|a\hat{\mathbf{p}}_i + \mathbf{b} - \mathbf{p}_i\|_1. \tag{A.3}$$

- *Scale-invariant depth map.* For depth map evaluation, the optimal scaling factor a^* is derived by minimizing the depth-weighted absolute difference:

$$a^* = \underset{a}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i - z_i|. \tag{A.4}$$

- *Affine-invariant depth map.* The optimal scale a^* and offset b^* for an affine transformation of the depth map are solved via:

$$(a^*, b^*) = \underset{a,b}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} |a\hat{z}_i + b - z_i|. \tag{A.5}$$

- *Affine-invariant disparity map.* In line with standard practices for disparity-based alignment [52], we perform a least-squares optimization in the disparity space. The scale a^* and shift b^* are determined by:

$$(a^*, b^*) = \underset{a,b}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} (a\hat{d}_i + b - d_i)^2, \tag{A.6}$$

  where $\hat{d}_i$ represents the predicted disparity and $d_i = 1/z_i$ represents the ground truth disparity. To avoid invalid depth values resulting from zero or negative disparities, the transformed disparity is lower-bounded by $1/z_{\text{max}}$ (the inverse of the maximum valid depth). The final aligned depth $\hat{z}_i^*$ is thus obtained by:

$$\hat{z}_i^* := \frac{1}{\max(a^*\hat{d}_i + b^*, 1/z_{\text{max}})}. \tag{A.7}$$

Metric Geometry. The following geometry related metrics are used in the evaluation:

- *Metric depth.* For metric depth evaluation, we directly assess the model's raw output against the ground truth without any alignment or value clipping, except when a method explicitly requires specific post-processing as part of its standard inference pipeline.
- *Metric point map.* To evaluate metric point maps, the predictions are only aligned via a global translation vector $\mathbf{b}^*$, which is calculated to minimize the depth-weighted spatial error:

$$\mathbf{b}^* = \underset{\mathbf{b}}{\operatorname{argmin}} \sum_{i \in \mathcal{M}} \frac{1}{z_i} \|\hat{\mathbf{p}}_i + \mathbf{b} - \mathbf{p}_i\|_1. \tag{A.8}$$

A.3 Evaluation Details

The methods presented in Table 1 and Table 2 of the main paper often employ disparate evaluation protocols, making a direct comparison challenging. To ensure a fair and consistent assessment across our diverse test set of perspective and panoramic data, we adopt a unified protocol for each camera format. Specifically, one standardized protocol is applied to all perspective datasets, and a separate one is used for all panoramic datasets.

For perspective datasets, as described in the main text, we adhere to the established MoGe protocol [72]. At inference time, images from various datasets with differing original resolutions are resized to a height of 518 pixels, while the width is scaled proportionally to maintain the aspect ratio. The resolution of the ground truth (GT) used for evaluation is consistent with the MoGe protocol, and our model's predictions are subsequently resized via interpolation to match the GT resolution before evaluation.

For panoramic datasets, we adhere to the DA^2 protocol [36]. At inference time, each panoramic (equirectangular) image is projected into six 518×518 perspective patches. The final 512×1024 depth prediction map is then generated by back-projecting the results from the cube faces to the equirectangular format. The GT resolution for evaluation aligns with the DA^2 protocol, and our predicted depths are interpolated to match this resolution.

B Complete Evaluation on Individual Datasets

In the main text of the paper, we have provided detailed evaluation results for panoramic datasets. However, due to space constraints, the performance on perspective image benchmarks was primarily reported as average metrics across multiple datasets to provide a macroscopic view of our model's capabilities. Evaluating a universal monocular depth estimation model necessitates a granular analysis across diverse scenes. Therefore, in this supplementary material, we provide the complete, unaggregated evaluation results specifically for perspective scenes.

Table A1 lists the performance of DepthMaster alongside state-of-the-art baselines on each individual perspective dataset used in our evaluation suite (such as NYUv2, KITTI, ETH3D, etc.). By presenting these granular results, we offer a more transparent and comprehensive understanding of the model's robustness and specific strengths across different domains in perspective depth estimation.

C Training and Validation Datasets

This section details the datasets used for training and evaluating DepthMaster. Our methodology aims to achieve robust zero-shot monocular depth estimation for both perspective and panoramic images. This is achieved by training the model on a large-scale, mixed-data regime and subsequently validating its generalization capabilities on a diverse suite of unseen evaluation benchmarks. These benchmarks were intentionally selected to encompass both perspective and panoramic formats, as well as diverse data distributions that span indoor and outdoor environments and comprise both real-world and synthetic scenes, to rigorously test the model's versatility. The specifics of our training and validation sets are detailed below.

C.1 Training Datasets

Table A2a summarizes the composition of our training set, including each dataset's domain, frame count, type, and the final calculated sampling weight. Our training dataset is a curated aggregation of 17 publicly available datasets from diverse domains, totaling approximately 4.0 million frames. This collection comprises one panoramic dataset, Structured3D [96], and 14 perspective datasets. To balance these heterogeneous sources during training, we employ a principled sampling strategy. Given that our training data includes only one panoramic dataset, we assign it a fixed sampling ratio of 20% to ensure its adequate representation.

The remaining 80% of the sampling budget is allocated to the 16 perspective datasets, including Hypersim [53], A2D2 [17], MVS-Synth [25], GTA-SfM [66], UrbanSyn [21], TartanAir [74], IRS [70], ARKitScenes [4], MegaDepth [38], BlendedMVS [87], KenBurns [46], Waymo [60], Argoverse2 [75], Taskonomy-tiny [95], CARD [14], and PointOdyssey [97]. The sampling ratio for each dataset is given in Table Table A2a.

C.2 Validation Datasets

To assess the zero-shot generalization capability of DepthMaster, we evaluate it on a comprehensive benchmark of 13 distinct datasets, as summarized in Table A2b. These datasets were not seen during training and were intentionally selected, including both perspective and panoramic benchmarks, allowing for a thorough evaluation of the model's flexibility. The evaluation sets are categorized as follows:

- **Perspective Datasets:** This category comprises ten datasets that cover a wide variety of scenes. It includes indoor environments (NYUv2 [57], iBims-1 [31], HAMMER [28]), outdoor driving and in-the-wild scenarios (KITTI [61], DDAD [18], Spring [45]), mixed indoor/outdoor scenes (ETH3D [54], DIODE [62]), object-centric views (GSO [12]), and synthetic sequences (Sintel [7]).
- **Panoramic Datasets:** This category assesses the model's performance on full 360-degree equirectangular images. It consists of three large-scale indoor datasets: the real-world Stanford2D3D [3] and Matterport3D [9], and the synthetic PanoSUNCG [63]. Due to the lack of established outdoor panoramic benchmarks, we provide additional qualitative comparisons for outdoor scenes in Figure A4 and on our interactive project page.

Table A1: Evaluation results of our method and competing methods on each dataset.

Method	NYUv2		KITTI		ETH3D		iBims-1		GSO		Sintel		DDAD		DIODE		Spring		HAMMER		*Avg.*	
	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Re↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑	Rel↓	δ_1↑
Metric point map																						
UniDepth V1	4.80	98.3	4.52	98.5	22.4	63.1	10.8	92.8	-	-	-	-	11.4	89.5	12.8	88.9	-	-	18.0	79.5	12.1	87.2
UniDepth V2	4.83	98.0	5.88	97.5	9.46	95.0	5.23	97.9	-	-	-	-	13.3	90.3	17.0	80.8	-	-	15.0	83.9	10.1	91.9
UniK3D	4.51	97.9	5.41	97.7	7.79	97.0	4.56	98.3	-	-	-	-	14.8	89.9	8.23	93.9	-	-	14.3	85.0	8.18	92.3
Depth Pro	6.13	97.3	11.1	85.3	21.2	64.9	6.89	96.9	-	-	-	-	22.6	61.3	13.5	81.8	-	-	14.5	86.0	13.7	81.9
MoGe 2	4.44	98.3	7.44	94.4	7.19	97.7	5.63	97.4	-	-	-	-	11.4	87.9	7.85	92.3	-	-	13.4	87.0	8.19	93.6
DepthMaster (Ours)	5.19	97.8	8.89	94.9	8.78	93.7	5.49	97.2	-	-	-	-	11.4	86.7	8.35	92.6	-	-	9.80	95.1	8.27	93.9
Metric depth map (wo/ GT intrinsics)																						
DA V1	10.5	94.9	11.6	94.5	40.2	24.0	12.9	81.8	-	-	-	-	34.5	44.7	58.0	16.2	-	-	54.8	27.3	31.8	54.8
DA V2	16.4	80.9	10.6	88.6	36.1	36.3	11.1	91.7	-	-	-	-	41.7	37.5	41.2	22.1	-	-	52.1	38.9	29.9	56.6
UniDepth V1	7.59	97.6	4.69	98.4	56.9	14.9	23.8	57.6	-	-	-	-	13.8	85.1	17.1	71.9	-	-	38.2	46.7	23.2	67.5
UniDepth V2	10.6	92.8	8.58	95.4	20.7	69.5	9.52	93.2	-	-	-	-	18.4	77.6	43.0	51.8	-	-	38.2	46.8	21.3	75.3
UniK3D	8.87	94.2	7.97	96.4	15.7	82.2	9.06	94.3	-	-	-	-	20.7	73.8	17.2	76.4	-	-	28.9	60.2	15.5	82.5
Depth Pro	10.7	91.9	23.5	38.3	38.5	32.8	15.9	81.5	-	-	-	-	33.4	35.3	31.9	37.7	-	-	39.1	63.0	27.6	54.4
MoGe 2	7.33	96.1	18.1	62.9	10.4	90.8	13.6	83.0	-	-	-	-	15.8	73.0	17.5	66.4	-	-	26.9	65.6	15.7	76.8
DepthMaster (Ours)	8.78	95.4	15.0	83.5	15.5	78.6	12.5	89.0	-	-	-	-	14.7	80.2	17.4	70.2	-	-	15.7	79.1	14.2	82.3
Scale-invariant point map																						
UniDepth V1	5.33	98.4	5.96	98.5	18.5	77.6	5.29	97.4	6.58	99.6	33.0	48.9	11.4	90.2	12.3	91.0	33.1	49.8	4.83	98.5	13.6	85.0
UniDepth V2	5.59	98.3	5.41	98.0	6.58	97.2	5.56	98.1	4.53	99.7	27.2	56.3	13.4	91.2	12.0	93.4	31.9	46.0	4.20	99.2	11.6	87.7
UniK3D	4.73	98.2	5.08	98.1	5.69	98.1	4.42	98.1	3.55	99.9	25.6	58.5	15.3	89.9	8.70	94.3	29.2	54.8	5.06	99.5	10.7	88.9
Depth Pro	5.04	97.7	10.6	95.1	11.2	92.0	5.84	97.1	4.94	99.8	26.9	63.9	15.8	81.0	8.52	91.6	28.1	60.5	6.82	98.7	12.4	87.7
MoGe	4.86	98.4	5.47	97.4	4.58	98.9	4.63	97.1	2.58	100	22.3	69.5	12.3	90.3	6.58	94.5	4.84	96.4	6.45	98.1	7.46	94.1
MoGe 2	3.94	98.3	8.27	97.5	5.45	98.6	5.34	98.3	2.55	100	23.1	66.8	11.0	90.7	8.42	93.7	31.1	42.4	8.77	98.4	10.8	88.5
DepthMaster (Ours)	4.02	97.9	5.60	96.8	4.56	98.8	4.34	98.1	2.84	100	18.8	72.2	10.5	89.0	5.97	95.2	19.7	73.3	4.85	99.6	8.12	92.1
Affine-invariant point map																						
UniDepth V1	3.93	98.4	4.29	98.6	12.2	89.6	4.65	98.0	2.99	99.8	28.5	58.4	10.3	90.5	8.56	90.9	29.6	58.5	4.15	98.7	10.9	88.1
UniDepth V2	3.66	98.4	4.75	98.0	4.35	98.4	4.05	98.1	2.91	99.9	17.9	76.5	12.0	90.8	7.45	92.4	25.1	66.9	3.45	99.4	8.56	91.9
UniK3D	3.41	98.3	4.47	98.0	3.85	98.7	3.37	98.3	1.91	100	18.1	76.9	13.7	89.2	5.71	94.8	23.8	69.2	3.44	99.5	8.18	92.3
Depth Pro	4.36	97.9	9.15	90.7	7.73	94.0	4.34	97.4	3.16	99.7	19.6	74.5	14.4	81.2	6.28	93.7	25.0	66.0	5.31	98.8	9.93	89.4
MoGe	3.68	98.3	4.86	97.2	3.57	99.0	3.61	97.3	1.14	100	16.8	77.8	10.5	91.4	4.37	96.4	4.45	96.4	3.88	98.1	5.69	95.2
MoGe 2	3.33	98.4	6.47	96.4	3.89	98.7	3.65	98.5	1.16	100	17.4	77.0	10.1	90.3	5.13	94.9	24.5	63.7	4.19	99.1	7.98	91.7
DepthMaster (Ours)	3.54	98.0	5.02	97.2	3.18	99.1	3.13	97.9	1.33	100	15.3	79.4	9.50	89.8	4.07	96.2	14.9	82.0	1.79	99.7	6.18	93.9
Local point map																						
MASt3R	-	-	-	-	5.54	95.3	6.19	95.0	-	-	11.4	87.9	8.58	91.8	8.75	90.9	-	-	-	-	8.09	92.2
UniDepth V1	-	-	-	-	8.61	92.6	5.92	96.0	-	-	13.4	84.3	8.18	92.0	9.95	90.0	-	-	-	-	9.21	91.0
UniDepth V2	-	-	-	-	3.99	97.4	4.02	97.3	-	-	9.35	92.2	8.18	92.4	6.15	95.3	-	-	-	-	6.34	94.9
UniK3D	-	-	-	-	3.84	97.4	3.92	97.2	-	-	9.12	92.4	8.55	92.3	5.81	95.3	-	-	-	-	6.25	94.9
Depth Pro	-	-	-	-	4.76	96.9	4.11	97.5	-	-	10.8	89.5	8.08	92.4	6.80	94.1	-	-	-	-	6.91	94.1
MoGe	-	-	-	-	3.21	98.1	4.16	96.8	-	-	8.63	92.7	6.74	94.3	4.78	96.3	-	-	-	-	5.50	95.6
MoGe 2	-	-	-	-	3.27	98.2	3.61	97.7	-	-	8.13	93.2	6.57	94.3	5.09	96.1	-	-	-	-	5.33	95.9
DepthMaster (Ours)	-	-	-	-	3.51	98.1	3.87	97.5	-	-	7.28	94.4	6.69	94.2	5.30	95.8	-	-	-	-	5.34	96.0
Scale-invariant depth map																						
MASt3R	5.37	96.0	6.24	94.5	5.68	95.5	5.58	95.2	3.72	99.1	26.3	63.7	13.5	81.5	8.37	89.4	32.2	53.5	5.50	96.5	11.2	86.5
DA V1	4.77	97.5	5.61	95.6	9.41	88.9	5.53	95.8	5.49	99.3	28.3	56.7	13.2	81.5	10.3	87.5	27.3	59.1	6.88	96.4	11.7	85.8
DA V2	5.03	97.3	7.23	93.7	6.12	95.5	4.32	97.9	4.38	99.3	23.0	65.2	14.7	78.0	7.95	90.0	28.0	61.1	5.92	97.7	10.7	87.6
Metric3D V2	4.69	97.4	4.00	98.5	3.84	98.5	4.23	97.7	2.46	99.9	20.7	69.8	7.41	94.6	3.29	98.4	24.4	64.4	4.19	99.1	7.92	91.8
UniDepth V1	3.86	98.4	3.73	98.6	5.67	97.0	4.79	97.4	4.18	99.7	28.3	58.8	10.1	90.5	6.83	92.8	29.2	59.3	4.19	98.4	10.1	89.1
UniDepth V2	3.65	98.4	4.24	98.0	3.23	98.9	3.45	98.1	3.16	99.7	23.1	65.3	11.0	91.5	5.92	94.1	24.9	65.1	3.48	99.1	8.61	90.8
UniK3D	3.36	98.4	4.08	98.1	3.33	98.7	3.42	98.0	2.21	99.9	21.4	67.5	11.8	91.4	5.22	94.7	24.6	65.1	3.17	99.5	8.26	91.1
Depth Pro	4.42	97.6	5.47	96.2	7.54	94.1	4.13	97.4	2.18	99.9	23.9	68.7	14.0	82.0	7.05	92.0	25.1	63.8	4.36	98.9	9.81	89.1
MoGe	3.44	98.4	4.25	97.8	3.36	98.9	3.46	97.0	1.47	100	19.3	73.4	9.17	90.5	4.89	94.7	4.63	96.4	3.77	98.1	5.77	94.5
MoGe 2	3.44	98.2	4.11	98.0	3.55	98.7	3.16	98.2	1.49	100	19.6	71.6	8.91	91.2	5.30	94.6	20.0	72.4	3.96	99.2	7.35	92.2
DepthMaster (Ours)	3.60	97.9	5.03	97.0	2.86	98.8	3.20	98.1	1.63	100	16.1	77.1	9.44	88.9	4.80	95.2	17.7	74.8	3.19	99.7	6.76	92.7
Affine-invariant depth																						
MASt3R	4.67	96.7	5.79	95.1	4.64	97.0	4.62	95.6	2.85	99.4	21.3	70.3	12.5	83.4	5.79	94.1	27.4	62.8	4.21	96.8	9.38	89.1
DA V1	3.82	98.3	5.04	96.4	6.23	95.2	4.23	97.3	1.98	100	20.1	71.8	11.3	86.1	6.75	92.6	22.4	68.9	5.77	97.3	8.76	90.4
DA V2	4.16	97.9	6.77	94.3	4.63	97.2	3.44	98.3	1.44	100	17.1	76.6	13.4	81.8	5.41	94.6	23.7	68.7	4.73	98.9	8.48	90.8
Metric3D V2	3.94	97.6	3.50	98.4	3.24	99.0	3.28	98.3	2.10	99.4	26.6	71.7	7.15	94.8	2.75	98.7	21.0	72.5	3.02	99.0	7.66	92.9
UniDepth V1	3.40	98.6	3.55	98.7	4.92	97.5	3.76	98.2	2.48	99.9	24.9	64.1	9.46	90.8	4.90	96.2	25.2	67.3	3.55	98.9	8.61	91.0
UniDepth V2	2.96	98.6	3.85	98.1	2.95	98.5	2.64	98.4	1.37	100	13.3	83.2	10.5	90.9	4.05	96.5	20.1	75.4	2.48	99.6	6.42	93.9
UniK3D	2.81	98.6	3.77	98.1	2.96	98.7	2.56	98.4	1.19	100	13.7	83.0	11.7	89.5	3.62	97.0	20.3	75.3	2.43	99.6	6.50	93.8
Depth Pro	3.67	98.2	5.12	96.8	4.97	96.4	3.23	98.3	1.46	100	15.8	80.1	12.6	84.1	4.66	95.6	21.7	70.5	3.30	99.6	7.65	92.0
MoGe	2.92	98.6	3.94	98.0	2.69	99.2	2.74	97.9	0.94	100	13.0	83.2	8.40	92.1	3.16	97.5	4.34	96.4	3.00	98.3	4.51	96.1
MoGe 2	2.89	98.6	3.75	98.1	2.80	99.1	2.36	98.8	0.94	100	13.3	82.5	8.26	92.5	3.14	97.4	15.9	81.2	2.85	99.3	5.62	94.8
DepthMaster (Ours)	3.15	98.3	4.43	97.2	2.37	99.3	2.51	98.3	1.09	100	12.5	83.1	8.67	90.5	2.86	97.6	13.1	84.5	1.10	99.7	5.18	94.9
Affine-invariant disparity																						
DA V1	4.20	98.4	5.40	97.0	4.68	98.2	4.18	97.6	1.54	100	20.2	77.6	12.7	86.9	5.69	95.7	22.2	72.5	5.56	98.0	8.63	92.2
DA V2	4.14	98.3	5.61	96.7	4.71	97.9	3.47	98.5	1.24	100	21.4	72.8	13.1	86.4	5.29	96.1	24.3	70.6	4.97	99.1	8.82	91.6
Metric3D V2	13.4	81.5	3.76	98.2	4.30	97.7	8.55	92.3	1.80	100	21.8	72.4	7.35	94.1	7.70	90.2	23.3	68.1	3.17	99.2	9.51	89.4
MASt3R	5.07	96.8	5.93	95.5	5.25	96.4	5.39	95.7	2.98	99.7	30.2	65.1	13.0	83.6	6.41	94.3	37.3	53.2	4.41	97.2	11.6	87.8
UniDepth V1	3.78	98.7	3.64	98.7	5.34	97.2	4.06	98.1	2.56	99.9	28.6	60.7	9.94	89.1	5.95	95.5	30.0	61.6	3.64	99.1	9.75	89.9
UniDepth V2	3.38	98.7	3.99	98.0	2.97	99.0	3.15	98.3	1.30	100	17.2	79.9	10.2	90.2	4.43	96.4	24.4	69.6	2.51	99.6	7.35	93.0
UniK3D	3.24	98.6	3.88	98.1	2.89	99.0	3.03	98.4	1.16	100	17.0	80.0	10.5	90.1	3.95	97.0	25.8	68.9	2.47	99.6	7.38	93.0
Depth Pro	4.21	98.1	5.10	97.0	4.94	96.7	3.74	98.2	1.49	100	17.4	79.1	11.7	87.1	4.84	96.4	27.5	64.5	3.31	99.6	8.42	91.7
MoGe	3.38	98.6	4.05	98.1	3.11	98.9	3.23	98.0	0.96	100	18.4	79.5	8.99	91.5	3.98	97.2	6.43	93.7	3.30	98.5	5.58	95.4
MoGe 2	3.35	98.6	3.92	98.1	3.21	98.9	2.85	98.7	0.96	100	18.0	78.7	8.69	92.1	4.03	97.2	18.7	76.6	2.90	99.5	6.66	93.8
DepthMaster (Ours)	3.63	98.4	4.71	97.0	2.95	98.8	3.05	98.2	1.11	100	19.2	80.2	9.24	90.1	3.67	97.3	18.9	77.9	1.47	99.7	6.80	93.8

Table A2: Training and Evaluation Datasets. Left: Training datasets. Right: Evaluation datasets.

(a) **Training Datasets.** SfM: Structure-from-Motion; MVS: Multi-View Stereo.

Name	Domain	#Frames	Acquisition	Weight
Hypersim	Indoor	75K	Synthetic	5.12%
A2D2	Outdoor/Driving	196K	LiDAR	0.82%
MVS-Synth	Outdoor/Driving	12K	Synthetic	1.23%
GTA-SfM	Outdoor/In-the-wild	19K	Synthetic	2.86%
UrbanSyn	Outdoor/Driving	7K	Synthetic	2.15%
TartanAir	In-the-wild	306K	Synthetic	5.12%
IRS	Indoor	101K	Synthetic	5.73%
ARKitScenes	Indoor	449K	SfM/MVS	6.80%
MegaDepth	Outdoor/In-the-wild	92K	SfM/MVS	5.73%
BlendedMVS	In-the-wild	115K	SfM/MVS	12.27%
KenBurns	In-the-wild	76K	Synthetic	1.64%
Waymo	Outdoor/Driving	788K	LiDAR	5.00%
Argoverse2	Outdoor/Driving	1.1M	LiDAR	5.00%
Taskonomy-tiny	Indoor	300K	SfM/MVS	10.42%
CARD	Outdoor/Driving	79K	LiDAR	4.11%
PointOdyssey	Outdoor	237K	Synthetic	6.00%
Structured3D	Indoor	35K	Synthetic	20.00%

(b) **Evaluation Datasets.** Grouped by image type: perspective (top), panoramic datasets (bottom).

Name	Domain	#Frames	Acquisition
NYUv2	Indoor	654	RGB-D
KITTI	Outdoor/Driving	652	LiDAR
ETH3D	Indoor/Outdoor	454	SfM/MVS
iBims-1	Indoor	100	RGB-D
GSO	Object-centric	1,030	Synthetic
Sintel	Synthetic/Various	1,064	Synthetic
DDAD	Outdoor/Driving	1,000	LiDAR
DIODE	Indoor/Outdoor	771	LiDAR
Spring	Outdoor/In-the-wild	1,000	Synthetic
HAMMER	Indoor	775	Synthetic
Stanford2D3D	Indoor	1413	SfM/MVS
Matterport3D	Indoor	2024	SfM/MVS
PanoSUNCG	Indoor	3944	Synthetic

This diverse suite of evaluation benchmarks provides a rigorous testbed for validating the robustness and versatility of our model across different camera intrinsics and real-world domains.

C.3 Real Data Refinement

Real-world training data often suffer from noise and incompleteness, which hamper the model's ability to accurately reconstruct fine-grained structures. Previous works [86, 73] have also noted this issue and validated its impact. However, they did not open-source their data processing methods. Therefore, we implement our own strategies: depth enhancement and flying pixel filtering. The primary challenges in real-world data include blurred object boundaries and severe flying pixels at depth discontinuities. To address these issues, we introduce two key data refinement strategies: depth enhancement and flying pixel filtering.

Depth Enhancement. To resolve the issue of blurred boundaries in real-world ground truth, we enhance the data using predictions from a model trained on synthetic datasets, which naturally possess perfectly sharp edges. Specifically, after obtaining the high-quality relative depth prediction z^{pred} from the synthetic-trained model, we apply a robust alignment pipeline to recover the metric scale. We first employ RANSAC-based linear regression separately on near and far regions to establish a global scale s and shift t, yielding an initial metric depth $z^{\text{init}} = s \cdot z^{\text{pred}} + t$. To address local inconsistencies, we then optimize a learnable local scale map $S \in \mathbb{R}^{H \times W}$ to refine the depth as $z^{\text{opt}} = S \odot z^{\text{init}}$. This is guided by an As-Rigid-As-Possible (ARAP) smoothness constraint and a point cloud L_1 loss against the noisy ground truth z^{gt}:

$$\mathcal{L} = \sum_{p \in \mathcal{V}} \left\| S(p)\mathbf{P}^{\text{init}}(p) - \mathbf{P}^{\text{gt}}(p) \right\|_1 + \lambda \sum_{p \in \mathcal{V}} \left| S(p) - \frac{1}{|\mathcal{N}(p)|} \sum_{q \in \mathcal{N}(p)} S(q) \right|, \tag{C.1}$$

where $\mathbf{P}(p)$ denotes the 3D coordinate unprojected from pixel p, $\mathcal{V}$ is the set of valid pixels, $\mathcal{N}(p)$ represents the local spatial neighborhood, and λ is the smoothness weight. Finally, we use edge-aware Poisson blending to seamlessly integrate the refined metric depth while preserving the sharp discontinuities of the relative prediction. As illustrated in Figure A1, this enhancement process yields depth maps with significantly sharper edges and fine-grained structures, while maintaining robust scale alignment.

Flying Pixel Filtering. To address the issue of flying pixels commonly encountered in real-world datasets, we apply an edge-based filtering technique. Specifically, we detect depth discontinuities by computing the relative depth difference between adjacent pixels. A pixel p is classified as an unreliable edge region and subsequently masked out if it satisfies the following condition:

$$\max_{q \in \mathcal{N}_4(p)} \frac{|z(p) - z(q)|}{\min(z(p), z(q))} > \tau, \tag{C.2}$$

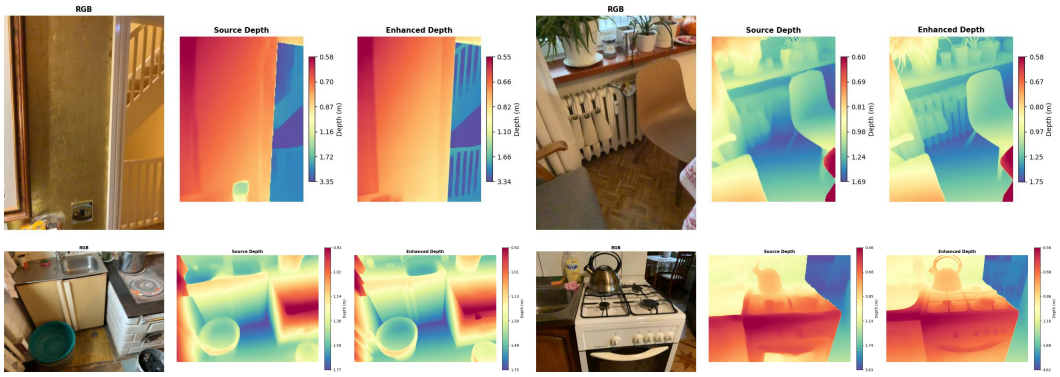


Figure A1: **Depth Enhancement on Real Data.** We enhance real-world depth maps using predictions from a synthetic-trained model followed by a scale alignment algorithm. The enhanced depth maps show significantly sharpened edges while preserving accurate scale alignment.

where $\mathcal{N}_4(p)$ denotes the 4-connected neighborhood of p, $z(p)$ is the depth value, and τ is a predefined relative tolerance threshold. This geometric strategy effectively eliminates the flying pixels commonly produced by LiDAR or SfM around object boundaries. Figure A2 demonstrates the depth maps and the corresponding 3D point clouds before and after filtering, showing a substantial reduction in flying pixels and cleaner geometric boundaries.

C.4 Training Data Scale Comparison

To further highlight the data efficiency of DepthMaster, we compare the training data scale of our method with representative state-of-the-art approaches in two complementary settings, as summarized in Table A3. On the perspective side (Table A3a), we report the total number of training frames used by each method for general monocular depth estimation. On the panoramic side (Table A3b), we report the number of panoramic frames used for 360° depth estimation.

DepthMaster achieves superior perspective accuracy with a smaller training corpus. As shown in Table A3a, our model uses only ∼4.0M frames, 2–17× fewer than existing baselines (DepthAnything V2 [86], Metric3D V2 [22], UniDepth V2 [50], UniK3D [48], MoGe [72], Depth Pro [6]). In particular, we deliberately exclude several large-scale real-world datasets adopted by MoGe, e.g., the Indoor subset of Taskonomy [95] (∼3.6M frames), yet our method still outperforms it on perspective benchmarks. This indicates that once the training set covers sufficiently diverse scenes, further scaling mainly adds annotation noise and scene redundancy, while our curated compact corpus is already sufficient for high-quality depth learning.

The advantage is more pronounced in the data-scarce panoramic domain, which is the central claim of DepthMaster. From Table A3b, we can see that our model uses only ∼35K panoramic frames from a single dataset (Structured3D [96]), yet still surpasses all panorama-oriented baselines: $4\times$ fewer than PanDA [8], $14\times$ fewer than DA360 [27], $17\times$ fewer than DA^2 [36], and over $54\times$ fewer than the SOTA metric panoramic method DAP [41] (∼1.9M). A particularly informative case is UniK3D [48]: under the same DINOv2 ViT-Large backbone and a comparable amount of panoramic data (∼39K), its panoramic accuracy still lags substantially behind ours (Tab. 1 in the main paper). Therefore, the panoramic gap is not determined by the amount of 360° data, but by how it is used.

We attribute this to the prior transferability unlocked by our framework. Instead of collecting massive 360° supervision to fit the distorted ERP representation, DepthMaster reuses the high-quality perspective priors from its backbone and curated perspective data, and transfers them to 360° inputs via our canonicalization design. A small panoramic set then suffices to adapt the model to the ERP distribution, while most geometric knowledge is inherited from the perspective side. This explains two patterns in Table A3b: (i) methods that purely scale up panoramic data (DA360, DA^2, DAP) exhibit clearly diminishing returns; and (ii) DepthMaster matches or exceeds their accuracy with one to two orders of magnitude less panoramic data, providing strong evidence that perspective-to-panoramic prior transfer is more effective than brute-force scaling.

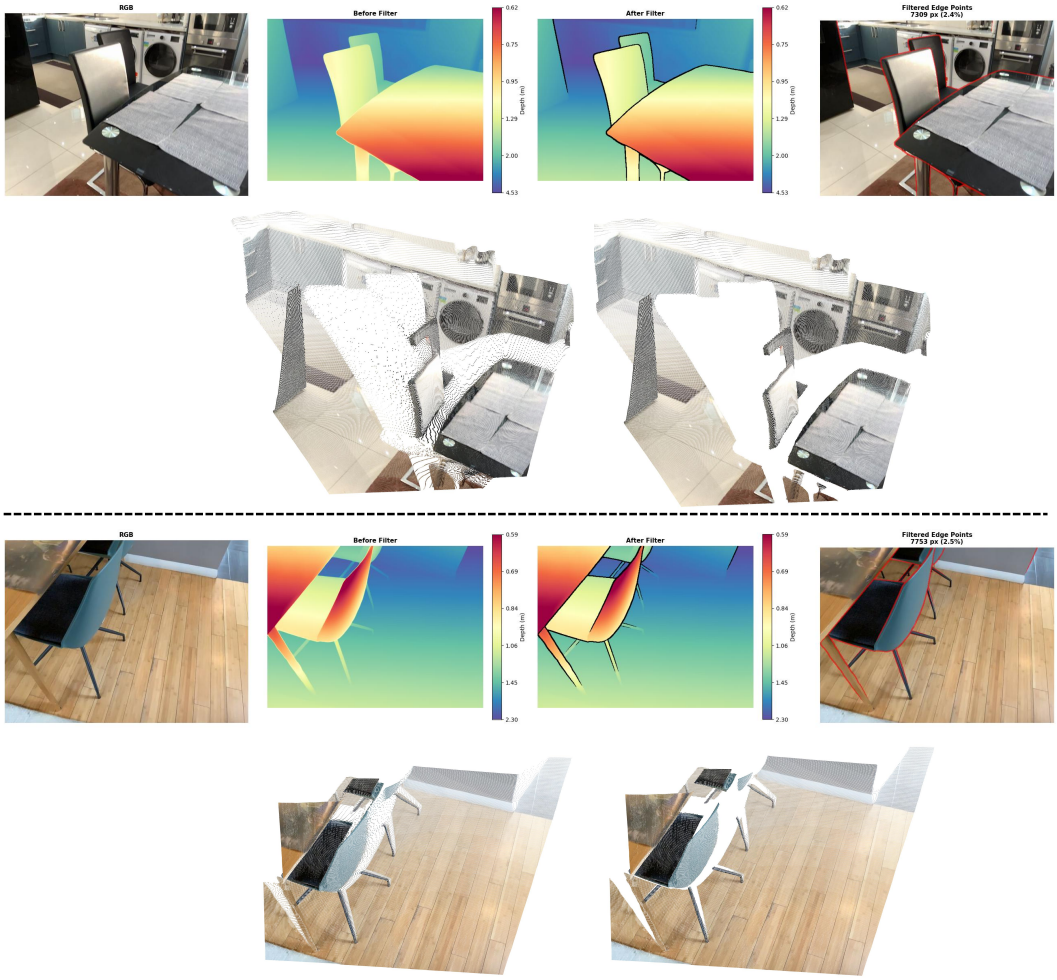


Figure A2: **Flying Pixel Filtering on Real Datasets.** We apply edge-based filtering to eliminate flying pixels at depth discontinuities. The visualizations of depth maps and 3D point clouds before and after filtering demonstrate a significant reduction in boundary artifacts.

Table A3: **Training data scale comparison.** All listed methods adopt DINOv2 as the backbone. Left: perspective-domain comparison (total training frames). Right: panoramic-domain comparison (panoramic training frames).

(a) **Perspective-domain comparison.**

Method	Backbone	#Training Frames
DepthAnything V2 [86]	DINOv2 ViT-Large	∼62M
Depth Pro [6]	DINOv2 ViT-Large	∼ 6M
Metric3D V2 [22]	DINOv2 ViT-Large	∼16M
UniDepth V2 [50]	DINOv2 ViT-Large	∼16M
MoGe [72]	DINOv2 ViT-Large	∼9M
UniK3D [48]	DINOv2 ViT-Large	∼10M
DepthMaster (Ours)	DINOv2 ViT-Large	**∼4.0M**

(b) **Panoramic-domain comparison.**

Method	Backbone	#Panoramic Frames
UniK3D [48]	DINOv2 ViT-Large	∼39K
PanDA [8]	DINOv2 ViT-Large	∼141K
DA360 [27]	DINOv2 ViT-Large	∼500K
DA^2 [36]	DINOv2 ViT-Large	∼606K
DAP [41]	DINOv3 ViT-Large	∼1.9M
DepthMaster (Ours)	DINOv2 ViT-Large	**∼35K**

D More Qualitative Comparisons

In this section, we present extensive qualitative comparison results to visually demonstrate the superior performance of DepthMaster on the universal monocular depth estimation task. As discussed in the main text, our model excels at generating depth maps characterized by high fidelity and structural consistency. To comprehensively evaluate its capabilities, we conduct a side-by-side comparison of

DepthMaster with representative models, including UniK3D, DA360, and DAP for panoramic scenes, as well as other state-of-the-art methods for perspective scenes. These comparisons are designed to highlight our model's advantages in recovering fine-grained details, preserving sharp boundaries, and handling complex geometric structures.

Panoramic Indoor Scenes. As shown in Figure A3, we provide a detailed comparison between our method and competing methods (UniK3D, DA360, DAP) on indoor panoramic scenes. The visual results include both depth maps and the corresponding 3D point clouds. DepthMaster demonstrates outstanding performance in preserving precise geometric structures. In contrast, the comparison methods frequently suffer from geometric inaccuracies, such as distorted walls and inconsistent structural layouts.

Panoramic Outdoor Scenes. Figure A4 presents a detailed comparison of our method against UniK3D, DA360, and DAP on outdoor panoramic scenes. The depth maps clearly show that our method produces much sharper edges. Furthermore, the point cloud visualizations demonstrate that DepthMaster preserves more fine details and maintains significantly better geometric structures across various scene styles. Additionally, our method features the unique capability to predict a sky mask, which the other methods cannot achieve. For point cloud visualization, we uniformly apply our predicted sky mask across all methods to mask out the sky regions.

Perspective Scenes. As illustrated in Figure A5, we additionally compare our method with state-of-the-art models, including Depth Pro, UniK3D, UniDepth, and MoGe on perspective images. The visualizations, which consist of both depth maps and 3D perspective views, demonstrate that DepthMaster possesses a robust capability to preserve fine-grained details and accurate geometric structures. While baseline methods often struggle with structural consistency, our approach produces highly reliable geometric predictions.

E Limitations and Future Directions

Following our discussions in the Section of Conclusion of the main paper, here we provide more analysis of DepthMaster's limitations, framing them as key directions for future research.

Our model shows limited generalization to highly abstract out-of-distribution (OOD) data. Although DepthMaster demonstrates strong generalization capabilities across various stylized scenes, it still struggles with highly abstract artistic styles, such as ink wash paintings. As illustrated in Figure A6, when confronted with such images, not only our DepthMaster but also all other state-of-the-art methods—including Depth Pro, UniK3D, UniDepth, and MoGe—fail to perceive the underlying 3D geometry, erroneously predicting them as flat planes. As a promising future direction, we plan to leverage image editing and style transfer models to synthesize a large-scale dataset featuring diverse abstract styles. Training on such constructed data will specifically enhance the model's depth perception capabilities in non-photorealistic and highly abstract domains.

F Broader Impacts

A unified depth estimation model that generalizes across perspective and panoramic cameras can benefit robotics, autonomous navigation, embodied AI, augmented reality, and large-scale 3D content creation by reducing the need for domain-specific retraining. To the best of our knowledge, this technology has no obvious negative societal impacts.

G Asset Licenses and Terms of Use

All training and evaluation datasets used in this work are obtained from their official project pages or authorized mirrors and are used for non-commercial academic research under their respective licenses or terms of use. We cite the original dataset papers throughout the main paper and supplement, and we do not redistribute any third-party data as part of this submission. For datasets with explicit public licenses, we follow those licenses directly (e.g., ARKitScenes under CC BY-NC-SA 4.0 and KITTI under CC BY-NC-SA 3.0); for datasets distributed through access agreements or project-specific terms (e.g., Structured3D, Matterport3D, A2D2, Waymo Open Dataset, and Argoverse 2), we follow the official download agreements and usage restrictions provided by the original creators.

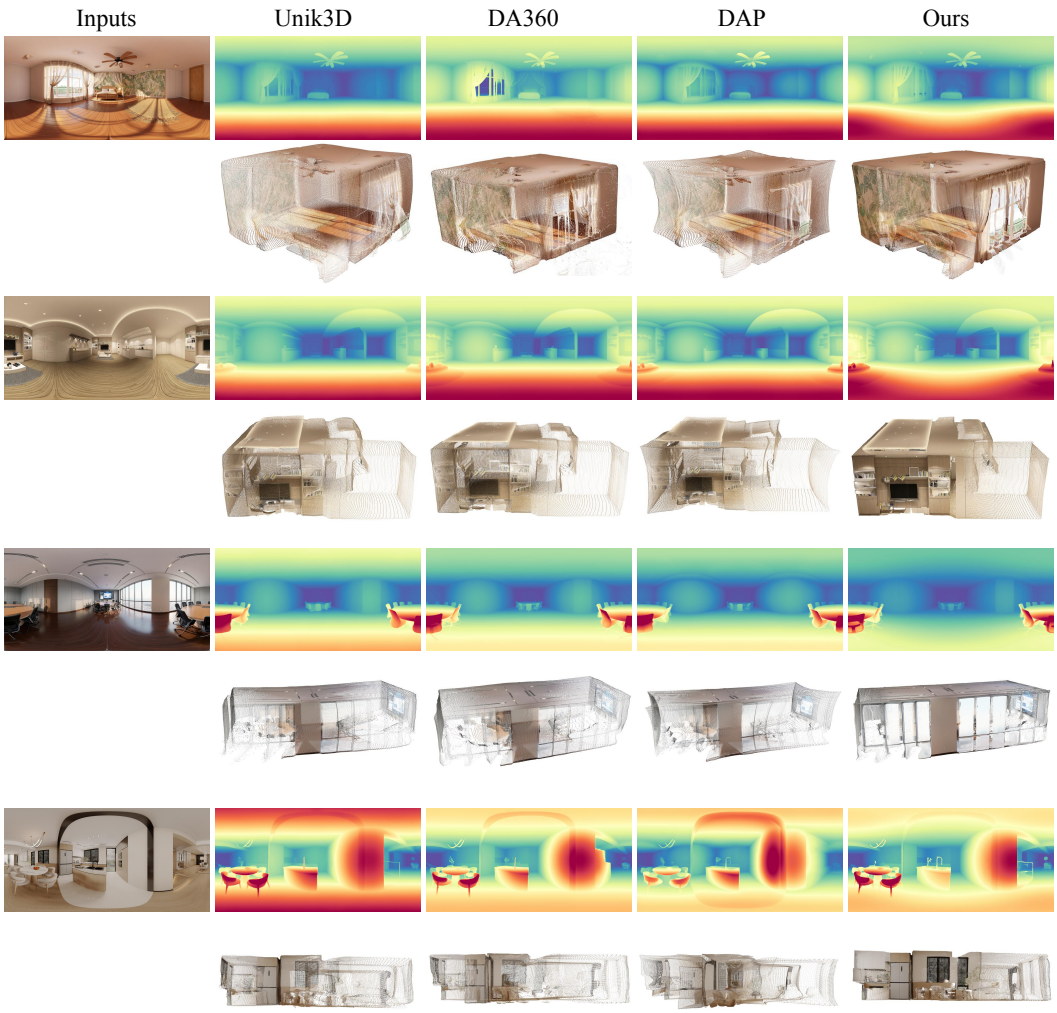


Figure A3: **Qualitative comparison on panoramic indoor scenes.** Compared with UniK3D, DA360, and DAP, DepthMaster achieves superior geometric accuracy and avoids issues such as wall distortions commonly seen in the baselines.

For baseline comparisons that require official implementations, we rely on the corresponding public codebases and checkpoints only through their official release channels and under their repository licenses or model terms. Any future release of our training or evaluation code will include references to these upstream assets and will require users to obtain third-party datasets or checkpoints directly from the original sources under their own licenses and terms.

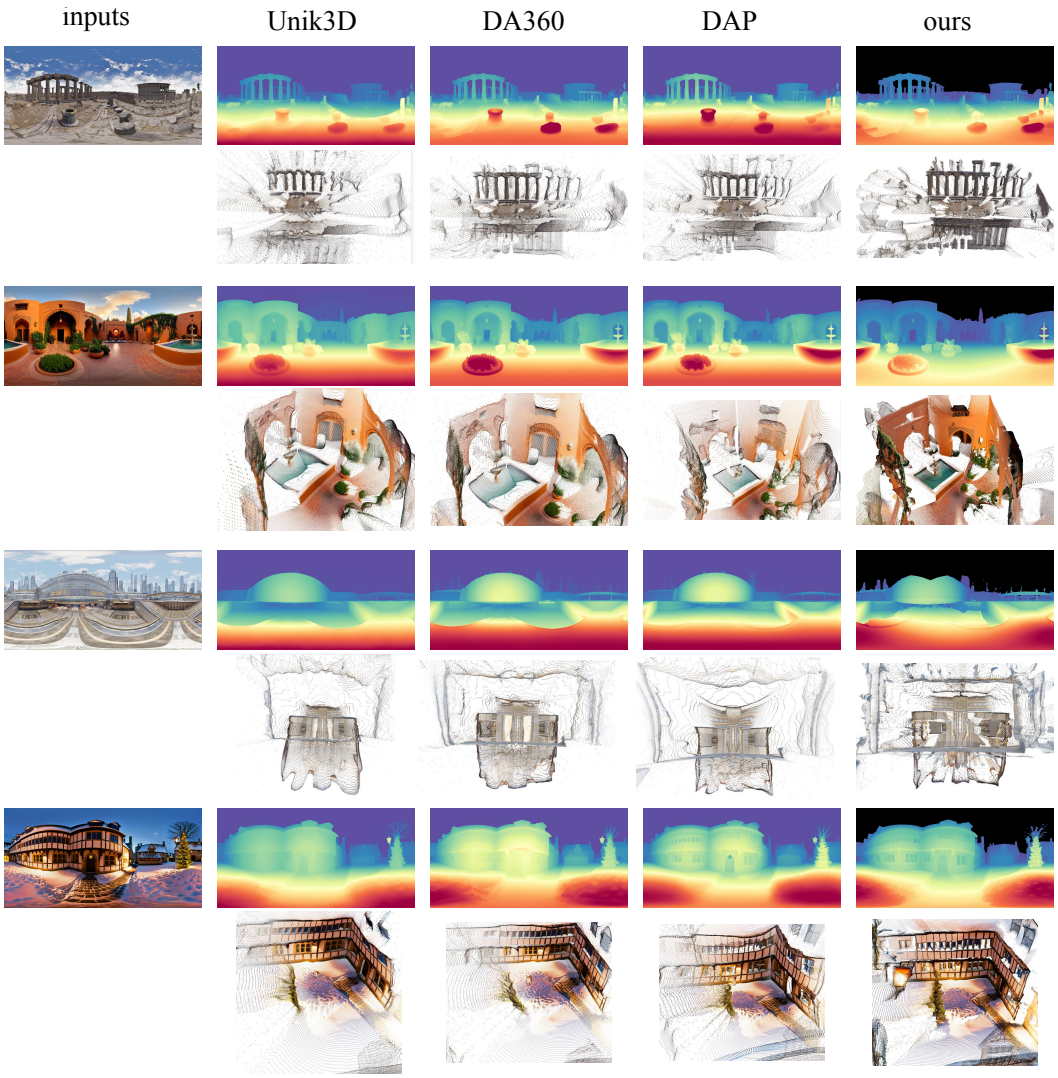


Figure A4: **Qualitative comparison on panoramic outdoor scenes.** DepthMaster yields sharper boundaries and better preserves geometric structures across different styles compared to UniK3D, DA360, and DAP. Note that all point clouds are visualized using the sky mask predicted by our method, as the baselines lack this capability.

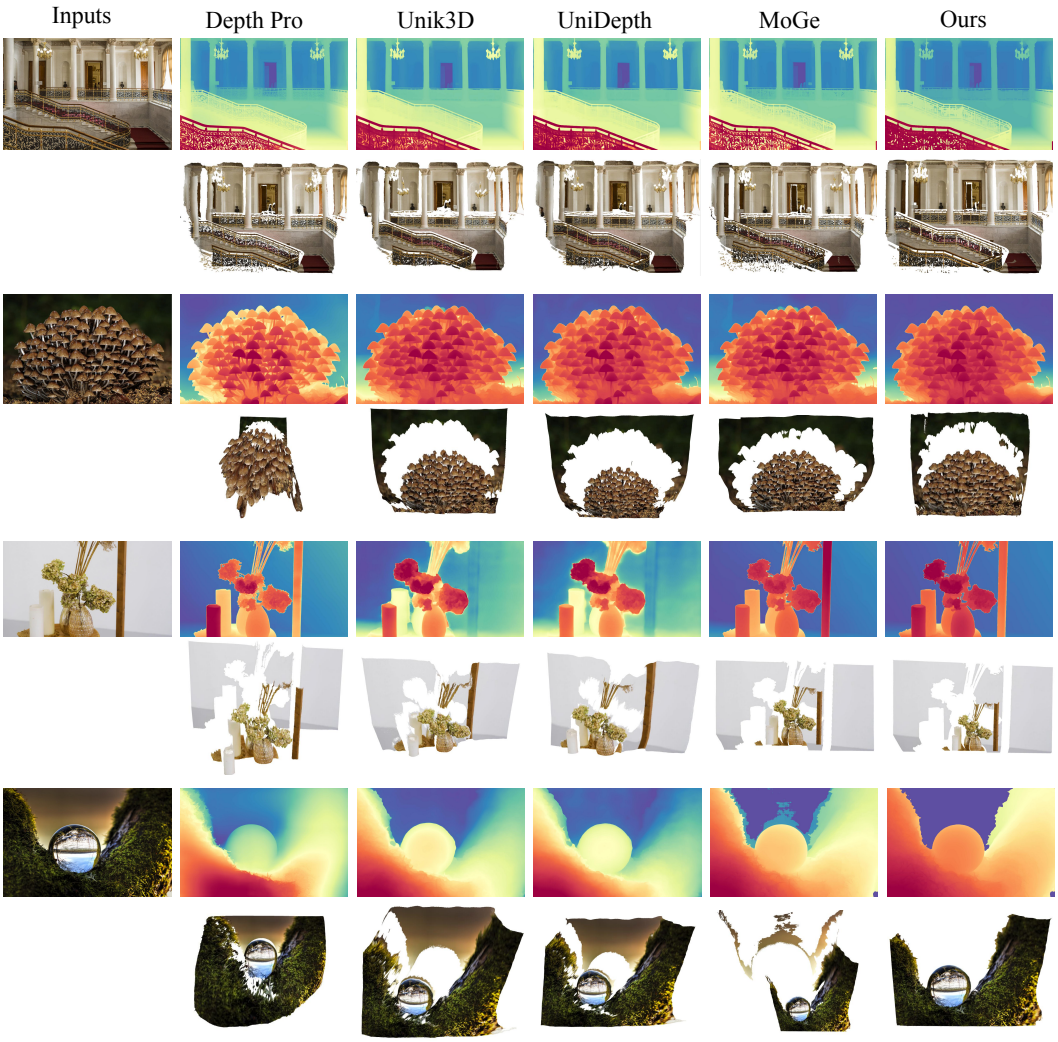


Figure A5: **Qualitative comparison on perspective scenes.** Compared with Depth Pro, UniK3D, UniDepth, and MoGe, DepthMaster demonstrates a strong capability in preserving fine-grained details and geometric structures, as evidenced by both the depth maps and 3D perspective views.



Figure A6: **Limitation on highly abstract artistic images.** For abstract styles such as ink wash paintings, our method, along with state-of-the-art methods (Depth Pro, UniK3D, UniDepth, and MoGe), fails to perceive the 3D geometry and incorrectly predicts the scenes as flat planes.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the paper's main contributions and scope, including the unified formulation, camera conditioning, CCL, and zero-shot evaluation on 13 datasets. These claims are supported by the main quantitative and qualitative results in Sections 1, 3, 4, and Appendix B.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: We discuss the limitations of the proposed method in the conclusion and further elaborate them in the appendix, including failure cases on highly abstract OOD artistic images and possible future directions. See the paragraph "Limitations" in the conclusion and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not present formal theoretical results such as theorems, lemmas, or proofs. Our contributions are methodological and empirical.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: We specify the model design, training recipe, data composition, sampling strategy, evaluation protocols, and full per-dataset quantitative results in the main paper and appendix. See Sections 3, 4 and Appendix A, A.3, B, and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: The paper states that the code and models will be released, but an anonymized code or model release with full instructions is not included in the current submission. Therefore, open access sufficient for faithful reproduction is not yet available at submission time.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The paper describes the training and evaluation settings, including datasets, protocols, optimizer, learning rate schedule, image resolution, iterations, augmentations, and benchmark definitions. See Section 4 and Appendix A, A.3, and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

Justification: The paper reports single-run quantitative results and ablations, but does not provide error bars, confidence intervals, or formal significance tests. We therefore answer No for this item.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: We disclose the training hardware, batch size, iteration budget, processed sample count, and measured inference latency and memory used for the reported experiments. See the "Prior Transferability and Efficiency" paragraph in Section 4 and Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: To the best of our knowledge, this research conforms to the NeurIPS Code of Ethics. The work is based on standard computer vision benchmarks and does not involve crowdsourcing or human-subject studies.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: We discuss the potential positive societal impacts of the proposed system and state that, to the best of our knowledge, it has no obvious negative societal impacts. See Section F in the supplemental material.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper studies monocular depth estimation rather than a high-risk generative model or a scraped dataset release. We therefore consider this question not applicable to the current work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: We cite the original third-party assets used in training and evaluation and summarize their licenses or official terms of use, together with our non-redistribution policy for upstream assets, in Section G of the supplemental material.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No].

Justification: The paper states that the code and models will be released, but no anonymized release package or accompanying documentation is provided in the current submission. Therefore, the answer is No at this stage.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A].

Justification: This work does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted on existing benchmark datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: This work does not involve human-subject research, so IRB approval or an equivalent review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs are not used as an important, original, or non-standard component of the core method in this research. Therefore, declaration is not required.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.