

Honey, I Shrunk the Arc de Triomphe!

Yuanbo Xiangli^{1,2}, Hanyu Chen¹, Xueqing Tsang¹, and Noah Snavely¹

¹ Cornell University, ² Shanghai Jiao Tong University

Abstract. Metric scale monocular geometry estimation has seen significant progress through large-scale data aggregation, yet current foundation models suffer from a persistent “scale-collapse” phenomenon: distant landmarks and vast landscapes are metrically underestimated. This performance gap stems from a training data bottleneck, where existing metric-scale datasets are hardware-constrained to unvaried street-level LiDAR or short-range indoor scans, or consist of synthetic data that lacks the semantic complexity of the physical world. To bridge this gap, we curate a new metrically-grounded, in-the-wild dataset that we call *MetricScenes*, gathered from a variety of sources including Internet photo collections and stereo imagery. We estimate camera poses and initial depth maps for each scene using off-the-shelf methods, and recover absolute scale from geo-tagged metadata as well as known stereo camera baselines. We also improve the quality of depth maps derived from *MetricScenes* via a new two-stage Poisson completion method. Fine-tuning MoGe-2 on our dataset significantly mitigates scale-collapse and achieves superior metric accuracy in unconstrained, open-domain scenes while maintaining state-of-the-art performance on standard benchmarks. Project page: <https://metricscenes.github.io/>.

Keywords: In-the-wild metric-scale dataset · Monocular metric geometry · Monocular metric depth

1 Introduction

Determining absolute scene scale from a single image is an inherently ill-posed task. Consider the tourist photo of the Arc de Triomphe in Fig. 1. From visual information alone, we can’t be completely certain that we’re looking at the monumental real-world landmark, as opposed to a cleverly constructed toy scene. However, there are visual cues to absolute scale, some of which humans rely on, such as reference objects of known scale (*e.g.*, people). Modern monocular geometry estimation (MGE) models ought to be able to perform a similar deduction via such semantic priors and other cues.

However, we found that current state-of-the-art methods frequently fail to estimate correct scene scale, leading to a persistent *scale-collapse* phenomenon in “in-the-wild” scenarios. Fig. 1 shows an example where clear semantic references (people) are present, yet where models like MoGe-2 [33] exhibit a significant scale inconsistency across the range of distances: the predicted metric scale for near-field objects is plausible—in this case, the tourists have a plausible height—yet the scale for far-field structures is dramatically underestimated—here, the Arc

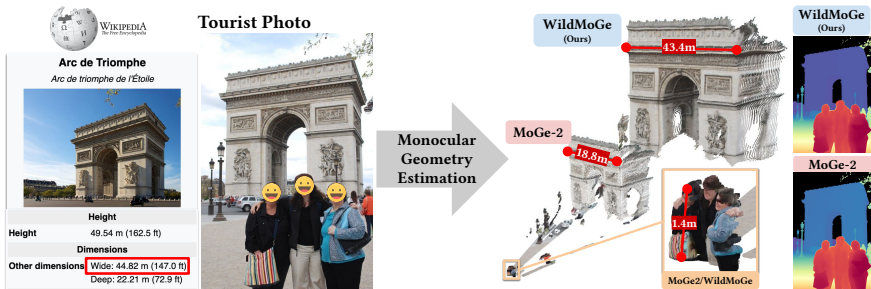


Fig. 1: Scale-collapse in metric geometry estimation. State-of-the-art monocular geometry estimation (MGE) methods like MoGe-2 [33] exhibit a persistent scale-collapse phenomenon, where far-field structures are metrically shrunk and pulled toward the camera. According to Wikipedia, the real Arc de Triomphe has a physical width of 44.8 m, but MoGe-2 predicts a metric point cloud of width just 18.8 m. By fine-tuning on our proposed MetricScenes dataset, our model, WildMoGe, recovers a more faithful landmark scale (43.4 m) while maintaining consistent metric estimation for foreground subjects (~ 1.4 m).

de Triomphe in the background is metrically predicted to be just 18.8 m wide, which is more than $2\times$ smaller than the ground truth width (44.8 m). MoGe-2 has posited a miniaturized landmark, despite cues to the contrary.

We argue that this performance gap is due to limitations in available training data. Current metric-scale, real-world datasets are restricted to two narrow categories due to hardware constraints: 1) terrestrial LiDAR datasets like KITTI [30] and Argoverse 2 [38], which are predominantly confined to vehicle-mounted street scenarios featuring limited diversity, and 2) active 3D scanner-captured data, such as ScanNet++ [43] and ARKitScenes [1], which are restricted by their short effective range to small-scale environments like indoor rooms. Synthetic datasets attempt to bridge this gap through diverse perspectives and motions [4, 8, 34], but fail to capture the true metric diversity of the real world due to the reliance on procedural rules and limited 3D assets.

To enrich the availability of real-world metric data, we introduce *MetricScenes*, a large-scale dataset curated from a variety of commonplace visual sources including Internet photo collections and stereo imagery. These sources provide the environmental and semantic variety missing from existing hardware-constrained datasets. We reconstruct camera viewpoints and initial depth maps via off-the-shelf methods, then recover absolute physical scale by leveraging geolocated landmark metadata and stereo camera baselines. Specifically, we aggregate data from AerialMegaDepth [32], MegaScenes [29], and Stereo4D [11], and develop pipelines to extract metric-scale depth maps in each case. We observe that depth maps derived from SfM and MVS often lack foreground transients. Naive Poisson completion fails here because the gradient guidance produced by monocular models also suffer from *scale-collapse*, causing the solver to unnaturally enlarge foreground objects or distort boundaries. To remedy this, we propose a two-stage edge-aware Poisson completion algorithm that decouples background metric alignment from foreground integration, ensuring sharp

boundaries and consistent metric scale throughout the view. We show that fine-tuning MoGe-2 on MetricScenes significantly improves metric accuracy for unconstrained, in-the-wild photos. Our model generalizes effectively to novel scenes while maintaining performance comparable to state-of-the-art methods on standard benchmarks.

2 Related Work

Scale Ambiguity. Recovering absolute dimensions from a single image involves inherent scale ambiguity. Despite advances, resolving true physical scale remains a central challenge for monocular metric depth and geometry estimation (MDE/MGE). Early works like ZoeDepth [2] and LeRes [45] attempted to bridge this gap using domain-specific metric heads or focal length decoupling. Recent foundation models have introduced more sophisticated alignment strategies. The Metric3D series [10, 44] transform images into a virtual camera space to resolve metric ambiguity across diverse data sources. UniDepth v2 [23] employs a self-prompting camera module, which simultaneously learns to predict camera embeddings. MoGe-2 [33] decouples structural geometry from global scale, utilizing DINOv2 global tokens for scale prediction.

However, these architectural innovations are often insufficient for robust “in-the-wild” generalization. As shown in Fig. 1, state-of-the-art models suffer from scale-collapse in unconstrained environments, often underestimating the dimensions of distant structures. This performance gap suggests that the primary bottleneck is no longer architectural, but rather a fundamental lack of scale-diverse, real-world training data.

Metric-scale Datasets. Real-world metric depth datasets relied on physical sensors like terrestrial LiDAR (*e.g.*, KITTI [30], Argoverse 2 [38], Waymo [28]) and active 3D scanners (*e.g.*, ScanNet++ [43], ARKitScenes [1]). While scale-accurate, these are restricted to driving corridors with limited diversity or small-scale indoor rooms and often suffer from spatial misalignment and sensor asynchrony. To bypass these hardware limits, synthetic datasets, like urban-focused MatrixCity [17] and Synscapes [39], indoor-oriented Hypersim [24], and aerial-view TartanAir [34], provide noise-free, dense labels across diverse domains. To further improve generalization, internet-scale collections like MegaDepth [19] and BlendedMVS [42] use structure from motion methods [25] to reconstruct scenes with wide semantic diversity, though they recover geometry only up to an unknown scale.

We analyzed standard training data and found a critical imbalance: large-scale in-the-wild SfM datasets offer semantic variety but lack metric scale. Metric supervision is currently split between LiDAR-based datasets, comprising 59% of outdoor metric frames but limited to driving scenarios; and the rest are synthetic data, which lacks real-world complexity. No existing large-scale real-world dataset simultaneously provides absolute scale and in-the-wild diversity. Our work addresses this gap by creating MetricScenes, a real-world metric-scale dataset that provides depth training data that is both metric and sharp.

3 Generating Metric-Scale 3D Data

To fill critical gaps and add diversity missing in existing real-world metric datasets, we leverage widely available visual sources, including Internet photo collections and stereo imagery. In Sec. 3.1, we describe how we process noisy Internet photo collections and recover physical scale using geo-tagged metadata. In Sec. 3.2, we cover how we acquire high-quality dense depth maps from stereo video sequences and recover metric scale using known camera baselines. The initial depth maps we obtain from off-the-shelf vision methods often contain holes and missing regions. In Sec. 3.3, we describe a new two-stage edge-aware depth completion strategy that preserves correct scene scale and geometry while enhancing depth sharpness.

3.1 Internet Photo Collections

We gather imagery from AerialMegaDepth [32] and MegaScenes [29], which consist of diverse Internet-sourced photography, including historical archives, tourist snapshots, and professional imagery. Specifically, AerialMegaDepth enhances MegaDepth [19] by jointly reconstructing landmarks with geo-tagged aerial views rendered from Google Earth. MegaScenes provides a vast collection of SfM reconstructions recovered at an arbitrary scale. We leverage geo-tagged images from online mapping services to anchor these reconstructions to absolute physical dimensions.

Robust SfM and MVS. For AerialMegaDepth, we use the SfM results provided in their dataset. For MegaScenes, we implement a pipeline addressing both structural errors and local depth inaccuracies. First, to resolve potential Doppelgänger issues [5], where visually similar but geographically distant 3D surfaces cause spurious correspondences that mislead the reconstruction, we follow [40] to obtain sparse reconstruction results with MAST3R-SfM [7] and the Doppelgänger classifier. For both AerialMegaDepth and MegaScenes, after MVS [26], on the obtained geometric depths, we apply a stability filtering strategy to identify and remove unstable pixels with high depth variance, and leverage MoGe-2’s predictions to filter out depth-bleeding regions where background depth eats away foreground depth [18, 19]. This depth refinement pipeline ensures the final labels are geometrically consistent and free from dynamic artifacts. Examples are shown in Fig. 2.

Metric-Scale Recovery from Geolocations. To establish a globally consistent metric coordinate system, we leverage geo-referenced imagery to anchor our 3D reconstructions, as illustrated in Fig. 2. AerialMegaDepth achieves metric scale by jointly reconstructing Internet photos with Google Earth renderings from specified, geolocated viewpoints. Since the rendered viewpoints are inherently metrically grounded, the resulting models inherit the absolute metric scale. Since MegaScenes lacks metric information, we leverage online mapping sites to obtain geo-tagged street view images:

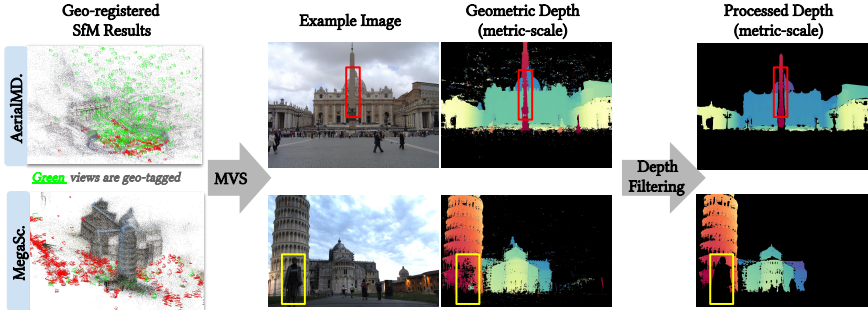


Fig. 2: Metric depth from Internet photo collections. Geo-tagged images obtained from online mapping sites can be used to scale SfM results to absolute metric scale. AerialMegaDepth [32] is reconstructed with pseudo-synthetic views rendered from Google Earth and scenes are scaled accordingly. MegaScenes [29] contains natively unscaled SfM results. We augment these SfM models with georeferenced street-level views to scale the geometry to physical dimensions. After scaling and running MVS, we apply a depth filtering method to remove transient objects (yellow box) and filter out depth-bleeding regions (red box).

- **Viewpoint sampling:** We ascertain the landmark’s outline and sample approximately 100 evenly distributed viewpoints along surrounding roads.
- **Coverage optimization:** To ensure full-facade coverage, we query for images with camera headings directed toward the scene center and apply geometric occlusion checks to remove obstructed views.
- **Data preference:** We prioritize imagery and metadata contributed by professional providers over those from individual users, as the former offers more accurate geolocation.

To register these geo-tagged images to existing reconstructed SfM models in MegaScenes, we use a method based on MAST3R. Specifically, MAST3R descriptors are used to compute matches between geo-tagged images and the existing reconstructed models. To resolve matching ambiguities arising from scene symmetries or repetitive elements, we use the Doppelganger classifier [40] to prune spurious correspondences in the scene graph. After registering the geo-tagged images, we use the COLMAP’s model aligner [25] to rotate the reconstruction into a Manhattan World frame, ensuring the dominant vertical and horizontal structures in the scenes are axis-aligned. To establish absolute metric scale and positioning, we apply RANSAC to robustly estimate a similarity transformation between the recovered 3D camera positions and their corresponding Earth-Centered, Earth-Fixed (ECEF) coordinates. To ensure robustness, we discard scenes with implausible scale factors, high mean registration errors, or low RANSAC inlier ratios.

3.2 Stereo Imagery

The Stereo4D dataset [11] provides a large-scale collection of diverse, real-world stereoscopic video sequences captured using VR180 devices. The dataset originally offers roughly-estimated metric scaled depth maps, using the assumption

that most VR180 stereo cameras have a baseline of approximately 6.3 cm (optimizing for the average human interpupillary distance). However, we observed that in practice, the physical baseline varies between different camera rigs. To ensure precise metric recovery, we utilize a subset of Stereo4D videos where camera configurations are explicitly documented in the video description.

Robust Multi-View Reconstruction. Stereo4D originally relies on an optical flow estimator, SEA-RAFT [36], to derive scene geometry. However, we find that imperfect stereo calibration often causes both optical flow and dedicated stereo matching algorithms (e.g., FoundationStereo [37]) to produce unreliable results. As illustrated in Fig. 3, the recovered geometry exhibits severe surface distortion, where parallel building facades unnaturally converge toward a distant point.

Therefore, to generate more reliable depth maps, we propose to use a *multi-view reconstruction* pipeline that re-estimates camera calibration, poses, and depth. For each video clip in Stereo4D, we first sample a target center frame and extract $N = 16$ surrounding frames with a temporal stride of $K = 3$. This spacing ensures sufficient temporal context and prevents the inclusion of many near-duplicate frames. Rather than relying on per-frame two-view stereo, we pass all $N + 1$ stereo pairs to a multi-view reconstruction model to jointly estimate camera poses and scene geometry. This approach leverages broader spatio-temporal context, leading to more accurate and consistent depth estimates than single-pair methods. We evaluated several methods, including π^3 [35], DepthAnything V3 [20] and MapAnything [13], and ultimately selected π^3 due to its geometric robustness and ability to recover sharp local details (see Fig. 3).

Metric-Scale Recovery from Camera Baselines. Since π^3 produces reconstructions in an arbitrary coordinate frame, we recover the absolute metric scale for each scene by aligning the average predicted baseline between the reconstructed stereo pairs to the known physical baseline. Specifically, denoting the physical baseline of the stereo rig as b_{gt} , we compute a global scale factor s as:

$$s = \frac{b_{\text{gt}}}{\frac{1}{N+1} \sum_{i=0}^N \|\mathbf{t}_{\text{L}}^{(i)} - \mathbf{t}_{\text{R}}^{(i)}\|_2}, \quad (1)$$

where $\mathbf{t}_{\text{L}}^{(i)}$ and $\mathbf{t}_{\text{R}}^{(i)}$ are the estimated translation vectors of the left and right cameras for the i -th sampled frame. Raw depth values d_{pred} are then scaled to produce the final metric depths: $d_{\text{metric}} = s \cdot d_{\text{pred}}$.

Frame Sampling and Filtering. Since frames within a clip are highly similar, we sample one sequence per clip and include only a single frame from the sequence in the final dataset to maximize dataset diversity. Specifically, we choose the *center-left frame* in the sequence as it benefits from the most bidirectional temporal context during the multi-view reconstruction process. To ensure high quality training data, we filter the target frame according to several criteria:

- **Image quality:** We discard underexposed frames based on their median intensity. Additionally, we filter out motion-blurred frames that lack sufficient high-frequency detail based on their Tenengrad gradient energy.

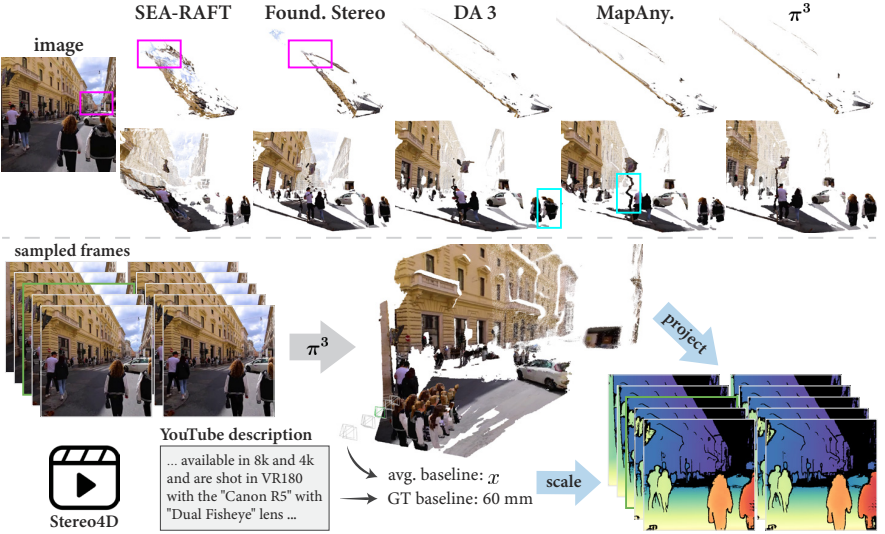


Fig. 3: Metric Depth from Stereo4D [11]. Top: Standard stereo matching [36, 37] often produces distorted geometry in poorly calibrated in-the-wild videos, as seen in the converging facades (magenta boxes). Among multi-view models [13, 20, 35], π^3 [35] maintains the most robust geometry and sharp local details (cyan boxes). Bottom: We process stereoscopic sequences via π^3 to obtain dense geometry and poses, then compute a global scale factor to align the predicted baseline with the camera’s physical baseline. This yields accurately scaled metric depth maps.

- **Geometric consistency:** To ensure geometric accuracy, we perform a depth reprojection check on the stereo pair and discard the frame if more than 10% of valid pixels are depth-inconsistent between the two views.
- **Calibration accuracy:** We reject frames violating expected stereo constraints. A frame is discarded if the inferred baseline deviates from the physical baseline by more than 10 mm (after scaling via Eq. 1) or if the relative camera rotation exceeds 1° .

To ensure reliable supervision from these pseudo-labels, we compute a robust scale factor between the derived metric depth from previous steps and the MoGe-2 prediction. If this factor exceeds $2\times$ or falls below $0.5\times$, we discard the frame as the scale is likely unreliable. This filtering is justified because π^3 consistently recovers near-field geometry (*e.g.* pedestrians close to camera, nearby buildings as in Fig. 3), a domain where MoGe-2 typically demonstrates high reliability.

3.3 Depth Fusion and Completion

Depth maps from prior stages are typically incomplete: Internet photos lack foregrounds and transient objects (Fig. 2), while stereo imagery lacks backgrounds due to low confidence on distant points (Fig. 3). Conversely, monocular models predict dense, continuous geometry but have problematic metric scale estimation. As a result, these sources become complementary: our processed depth

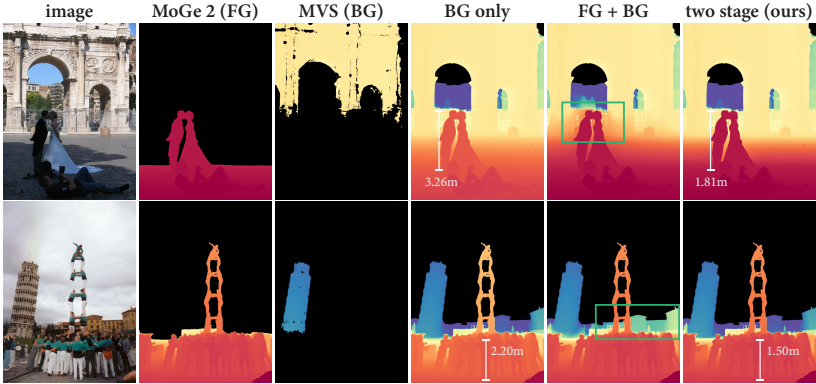


Fig. 4: Visual comparison of depth completion methods. With gradient guidance from MoGe-2, using only background anchors (BG only) preserves structure but uniformly mis-scales the entire map, yielding an implausible foreground (e.g., a 3.26m tall person). Combining foreground and background anchors in a single stage (FG+BG) forces the solver to reconcile conflicting scales, causing depth bleeding and scale drift (green boxes). Our two-stage approach maintains accurate metric scale for both foreground and background while ensuring sharp, artifact-free boundaries.

maps provide sparse metric anchors for scale, while monocular predictions offer structural guidance to fill holes and refine edges. Fusing them should produce depth maps that are both metrically accurate and visually complete.

Single-Pass Geometry Completion. To reconstruct missing depth, MoGe-2 introduced a *logarithmic-space Poisson completion* strategy. Their original approach integrates the incomplete ground truth depth map with guidance from a monocular depth model trained exclusively on synthetic data. Specifically, this is achieved by minimizing the gradient differences in the log domain:

$$\min \sum_{i \in \Omega} \|\nabla(\log d_i^*) - \nabla(\log \hat{d}_i)\|^2, \quad \text{s.t.} \quad d_i^* = d_i, \quad \forall i \in \partial\Omega, \quad (2)$$

where d_i^* is the solved depth, $\hat{d}_i$ is the predicted depth providing gradient guidance, and d_i represent fixed ground truth anchors bordering missing regions.

We found that this strategy effectively completes Stereo4D depth maps but faces a critical limitation on AerialMegaDepth and MegaScenes due to the scale-collapse phenomenon. Specifically, MoGe-2 often predicts an incorrect scale for distant landmarks, while maintaining a plausible scale for foreground subjects (Fig. 1). Because the Poisson solver treats the depth map as a continuous field, a scale correction forced by background anchors propagates globally through these gradient constraints. This results in a proportionate expansion of foreground objects, e.g. the physically impossible 3.26 m tall person in Fig. 4 (BG only), making foreground scale unreliable.

To resolve this scale conflict, we propose anchoring the Poisson completion by trusting MoGe-2 to recover the metric scale of foreground content while relying on filtered MVS results for the absolute scale of background regions. A naive solution would be to adapt Eq. 2 to use both the MVS background and MoGe-2

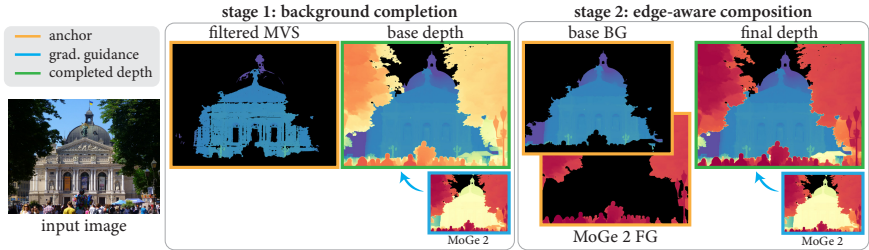


Fig. 5: Overview of the two-stage depth completion pipeline. Stage 1 performs a Poisson solve using the gradient guidance of MoGe-2’s depth map, constrained by the filtered MVS results. This yields a base depth map with an accurate background scale, though the foreground exhibits scale-drift. In Stage 2, the background isolated from the base depth map (base BG) and the MoGe-2 foreground (MoGe-2 FG) serve as joint anchors for the final completion. We perform an edge-weighted Poisson solve using these anchors, reusing MoGe-2’s gradient guidance to produce a globally consistent metric depth map with sharp local details.

foreground as fixed anchors. However, this approach fails because the solver is still forced to reconcile conflicting scales across the entire scene. As shown in Fig. 4 (FG+BG), unanchored transition regions such as a building’s base are inevitably warped toward the foreground to satisfy global gradient constraints.

Two-stage Edge-Aware Completion. We attribute geometric distortion in single-stage solves to anchor sparsity, which provides insufficient constraints to override the scale-collapse inherent in the gradients of MoGe-2’s depth prediction. Without a dense reference, the solver cannot reconcile the sparse metric anchors with a gradient field that pulls the background forward. This forces transition regions to warp as the geometry ‘stretches’ to satisfy collapsed gradients between the fixed foreground and background metric anchors. To address this, we propose a two-stage completion pipeline:

- **Stage 1: Background Completion.** We first apply Eq. 2 using only sparse background depth as anchors, yielding results like in Fig. 4 (BG only). In this stage, the background is reconstructed at a consistent metric scale without being warped by foreground depth constraints. We then discard the incorrectly scaled foreground regions, retaining only the completed background for the subsequent stage.
- **Stage 2: Edge-Aware Composition.** We combine the completed background from Stage 1 with the MoGe-2 predicted foreground to form a composite anchor. To preserve sharp boundaries and prevent depth-bleeding artifacts, the second Poisson solve utilizes an *edge-weighted* objective:

$$\min \sum_{i \in \Omega} \sum_{j \in \mathcal{N}(i)} w_{ij} \|(\log d_i^* - \log d_j^*) - (\log \hat{d}_i - \log \hat{d}_j)\|^2, \quad (3)$$

where the weight w_{ij} approaches zero across sharp depth discontinuities.

Example results are shown in Fig. 4 (last column). Because the background is now densely anchored by the Stage 1 output, it remains fixed to its correct

metric scale, while the edge weights ensure the foreground is integrated with crisp, artifact-free boundaries. Our full pipeline is illustrated in Fig. 5.

4 Evaluation

4.1 Dataset Overview

The MetricScenes dataset contains 47,579 images from 134 scenes in Aerial-MegaDepth (only real-world images), 29,583 images across 356 scenes from MegaScenes, and 22,549 frames from 1,725 videos¹ in Stereo4D. This diverse collection provides extensive coverage across a variety of perspectives and environments, including ground-level and aerial views, urban and natural landscapes, and both indoor and outdoor settings. We reserve a held-out set of 10 scenes from AerialMegaDepth, 10 scenes from MegaScenes, and 10 videos from Stereo4D to serve as validation and test sets.

4.2 WildMoGe

We fine-tune the MoGe-2 ViT-Large-Normal model on our MetricScenes dataset for 10,000 iterations with a batch size of 32 (approximately 3 epochs), ensuring model convergence while preventing significant divergence from the pre-trained geometric priors. We adopt the image cropping and data augmentation strategies from the original MoGe-2 implementation. We use a learning rate of 1×10^{-6} for the backbone and 1×10^{-5} for the remaining parameters. Further details can be found in the supplemental material.

4.3 Qualitative Results

In this section, we compare depth reconstructions of novel in-the-wild scenes obtained by our fine-tuned model, WildMoGe, as well as (vanilla) MoGe-2 [33], DepthAnything v3 [20], Metric3D v2 [10], UniDepth v2 [23], and DepthPro [3]. These results are visualized in Fig. 6. Additionally, we provide a comparison between WildMoGe and MoGe-2 on scenes prevalent in the standard training datasets, with results shown in Fig. 7.

From Fig. 6, we observe that WildMoGe consistently recovers more accurate absolute scales across diverse landmarks, closely matching ground-truth dimensions (*e.g.*, 31.4m vs. 32.4m for the Philadelphia Museum of Art, 46.7m vs 46.5m for Piazza della Signoria). MoGe-2, DepthAnything v3 and Metric3D v2 exhibit scale-collapse behavior, consistently underestimating the size of far-field structures. UniDepth v2 produces more realistic scales but still deviates from ground truth, and DepthPro often fails to recover absolute scale, producing results that are orders of magnitude smaller than reality. Note that these scenes are absent from the training set. This performance demonstrates that WildMoGe can generalize to unseen content, as opposed to simply memorizing training scenes.

¹ There are 1,725 videos, separated into 22,549 clips, and we sample 1 frame per clip.



Fig. 6: Metrology of novel in-the-wild scenes. The first column shows images with measurements obtained via Google Map’s measuring tool. We merge WildMoGe and MoGe-2’s results into a single column to highlight the accurate scaling achieved by our training scheme. WildMoGe consistently recovers more accurate absolute scales across diverse landmarks, whereas MoGe-2 [33], DepthAnything v3 [20] and Metric3D v2 [10] exhibit scale-collapse, underestimating the scale of background structures. While Unidepth v2 [23] produces more realistic scales, they still deviate from ground truth. DepthPro [3] often produces scales orders of magnitude smaller than reality.

Fig. 7 demonstrates that WildMoGe provides scale estimates consistent with MoGe-2 in common indoor and street environments. For instance, both models estimate a door height of 2.1m and a car length of approximately 3.2m to 3.3m. However, on the ETH3D [27] courtyard scene, WildMoGe recovers a more accurate, smaller scale than MoGe-2. Specifically, WildMoGe estimates a desk leg height of 71.6cm, closely matching the 72cm ground truth, whereas MoGe-2 overestimates it at 81cm. This indicates that WildMoGe is *not* merely biased toward predicting larger scales as in Fig. 1& 6, but is instead performing reliable estimation grounded in absolute metric geometry.

4.4 Quantitative Evaluation

Obtaining precise and dense ground-truth data for unconstrained, in-the-wild scenes is inherently challenging. Therefore, we provide quantitative results on our curated test set while acknowledging that these results may be less definitive due to potential imperfections in the test set labels. To provide a more comprehensive assessment, we also conduct evaluations on the standard benchmarks

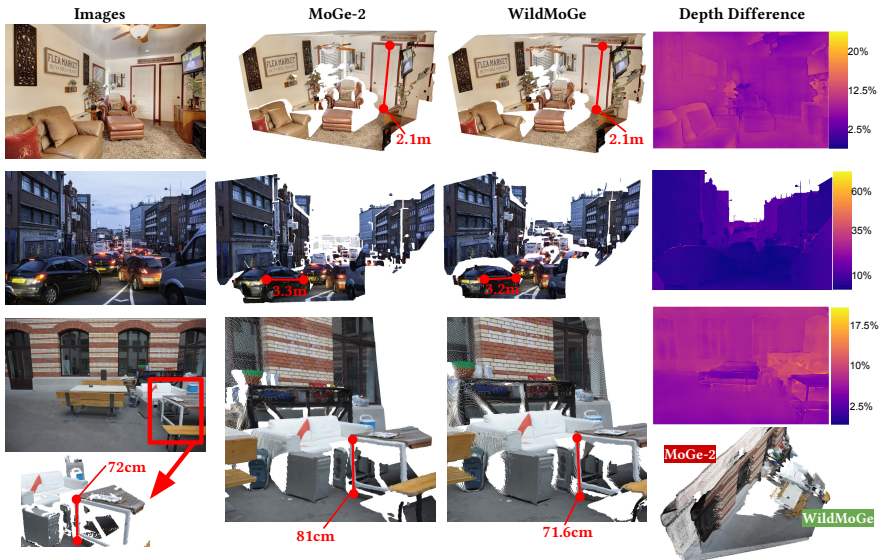


Fig. 7: Comparison on the standard scenes. We compare WildMoGe against MoGe-2 [33] on representative indoor and street-level scenes. In standard indoor and street contexts (Rows 1 & 2), WildMoGe provides scale estimates consistent with MoGe-2. On the ETH3D [27] courtyard scene (Row 3), WildMoGe achieves better accuracy, recovering a desk leg height of 71.6cm compared to the 72cm ground truth. This implies that WildMoGe’s performance is driven by precise metric grounding rather than a bias toward larger scales.

to demonstrate that our model is comparable to state-of-the-art performance in specialized environments while effectively bridging the gap between hardware-constrained training data and unconstrained, in-the-wild scenarios.

Benchmark. For the MetricScenes test set, we treat the partial dense depth maps generated *before* the Poisson completion step (Sec. 3.3) as our ground truth. For Internet photo collections, these labels are filtered geometric depth maps from MVS that have been metrically scaled using geo-tagged metadata (Sec. 3.1). For stereo imagery, the labels are from π^3 multi-view reconstruction results scaled by camera baselines (Sec. 3.2). We also follow MoGe-2 [33] to evaluate the accuracy of our model on 10 standard datasets: NYUv2 [22], KITTI [30], ETH3D [27], iBims-1 [14, 15], GSO [6], Sintel [4], DDAD [9], DIODE [31], Spring [21], and HAMMER [12].

Baselines. We compare our method to MoGe-2 [33], UniDepth V2 [23], Depth Pro [3], MAST3R [16], Depth Anything V2 [41] & V3 [20], ZoeDepth [2] and Metric3D V2 [10].

Relative Geometry and Depth. Following MoGe-2, we evaluate relative geometry to assess how effectively each method reconstructs scene structure from a single input image. We measure performance across multiple representations, including scale-invariant and affine-invariant point maps, local point maps within each detected objects, and scale-invariant depth, as well as affine-invariant

Table 1: Quantitative evaluation of relative and metric geometry. The top section evaluates on the standard benchmarks, while the bottom section evaluates on MetricScenes test set. Metrics are color-coded: red (best) and yellow (second best).

Method	Relative Geometry								Metric Geometry					
	Point			Depth			Point			Depth				
	Scale-inv. RelP↓	Affine-inv. δ ₁ ↑	Local RelP↓	Scale-inv. Rel ^d ↓	Affine-inv. Rel ^d ↓	Affine-inv. (disparity) δ ₁ ↑	(w/o GT Cam) RelP↓	(w/o GT Cam) δ ₁ ↑	(w/o GT Cam) Rel ^d ↓	(w/o GT Cam) δ ₁ ↑	(w/o GT Cam) Rel ^d ↓	(w/o GT Cam) δ ₁ ↑	(w/o GT Cam) Rel ^d ↓	(w/o GT Cam) δ ₁ ↑
<i>Evaluation on standard benchmarks</i>														
ZoeDepth	-	-	-	-	12.7	83.9	10.1	88.5	11.1	88.3	-	-	39.3	49.9
Metric3D V2	-	-	-	-	7.92	91.8	7.66	92.9	9.51	89.4	-	-	-	18.3
DA V2	-	-	-	-	10.7	87.6	8.48	90.8	8.82	91.6	-	-	29.9	56.6
DA V3	12.0	86.5	9.42	88.9	7.18	93.4	8.99	89.4	7.40	91.8	8.63	91.4	9.71	89.3
MASt3R	14.5	82.1	11.6	86.0	8.09	92.2	11.2	86.5	9.38	89.1	11.6	87.8	26.2	55.3
UniDepth V2	11.6	87.7	8.56	91.9	6.34	94.9	8.61	90.8	6.42	93.9	7.35	93.0	10.1	91.9
Depth Pro	12.4	87.7	9.93	89.4	6.91	94.1	9.81	89.1	7.65	92.0	8.42	91.7	13.7	81.9
MoGe-2	10.8	88.5	7.98	91.7	5.33	95.9	7.35	92.2	5.62	94.8	6.66	93.8	8.19	93.6
WildMoGe	10.1	89.3	7.68	92.2	5.50	95.8	7.49	91.8	5.63	94.7	6.97	93.7	8.39	93.1
<i>Evaluation on MetricScenes test set</i>														
MoGe-2	6.76	95.0	4.56	96.9	-	-	4.85	95.4	3.42	97.4	4.61	96.3	11.7	87.3
WildMoGe	5.24	97.0	3.67	97.9	-	-	4.02	97.1	2.98	98.2	4.15	97.1	7.59	93.4

depth and disparity. Specifically, the relative error for point maps is defined as $Rel^p = \|\hat{\mathbf{p}} - \mathbf{p}\|_2 / \|\mathbf{p}\|_2$ and for depth maps as $Rel^d = |\hat{z} - z|/z$. We also report the percentage of inliers using the thresholds $\delta_1^p = (\|\hat{\mathbf{p}} - \mathbf{p}\|_2 / \|\mathbf{p}\|_2 < 0.25)$ and $\delta_1^d = \max(\hat{d}/d, d/\hat{d}) < 1.25$ across the 10 evaluation datasets. ZoeDepth, DepthAnything V2, and Metric3D V2 are excluded from the point-based evaluations as these models do not provide camera intrinsic predictions. On the MetricScenes test set, we omit local point map evaluation because our ground truth depth maps can be sparse. Tab. 1 (left side) presents quantitative results on relative geometry and depth prediction. On the standard benchmarks, our WildMoGe achieves performance comparable to MoGe-2 while outperforming all other baselines. On the MetricScenes test set, WildMoGe consistently surpasses MoGe-2 across all metrics. This indicates that fine-tuning on MetricScenes does not severely compromise the model’s existing geometric priors; instead, it bridges the gap between near-field and far-field scale estimation on in-the-wild scenes, allowing the model to generalize effectively to unconstrained environments.

Metric Geometry and Depth. We evaluate the accuracy of metric-scale geometry and depth across seven datasets with metric-scale annotations: NYUv2 [22], KITTI [30], ETH3D [27], iBims-1 [14, 15], DDAD [9], DIODE [31], and HAMMER [12]. We report relative point error (Rel^p) and percentage of inliers (δ_1^p) on predicted metric point maps. Similarly, we evaluate metric depth accuracy via relative depth error (Rel^d) and depth inlier percentage (δ_1^d). Additionally, for methods that accept external camera parameters, we evaluate metric depth estimation using ground-truth camera intrinsics to isolate depth accuracy from errors in intrinsic parameter estimation. Results are in Tab. 1 (right side). While slightly worse than MoGe-2 on these metrics, WildMoGe outperforms most of the other baselines. This slight performance drop can be ascribed to the domain shift, and could be remedied by joint training. On the MetricScenes test set, WildMoGe outperforms MoGe-2 across all metrics. This suggests that while there is a trade-off on legacy datasets due to domain shift, the model successfully overcomes the scale-collapse phenomenon to achieve superior accuracy in unconstrained, in-the-wild environments.

Table 2: Ablation of finetuning MoGe-2 using different subsets of MetricScenes. Evaluation is conducted on the standard benchmarks and MetricScenes’s test set. We evaluate performance across relative geometry, metric geometry, and boundary sharpness. Metrics are color-coded: **red** (best) and **yellow** (second best). Boundary sharpness is reported on a subset of the standard benchmarks [4, 12, 15, 21], but is omitted for the MetricScenes test set.

Method	Relative Geometry										Metric Geometry						Sharpness Boundary		
	Point			Depth			Point			Depth									
	Scale-inv.	Affine-inv.	Local	Scale-inv.	Affine-inv.	Affine-inv. (disp)	(w/o GT Cam)	(w/o GT Cam)	(w/ GT Cam)	(w/o GT Cam)	Affine-inv.	(w/ GT Cam)							
	Rel $\downarrow$	$\delta^p\uparrow$	$\delta^d\uparrow$	Rel $\downarrow$	$\delta^p\uparrow$	$\delta^d\uparrow$	Rel $\downarrow$	$\delta^p\uparrow$	$\delta^d\uparrow$	Rel $\downarrow$	$\delta^p\uparrow$	$\delta^d\uparrow$							
Evaluated on the standard Benchmark														$F_{\uparrow}$					
All	10.1	89.3	7.68	92.2	5.50	95.8	7.49	91.8	5.63	94.7	6.97	93.7	8.39	93.1	15.6	83.4	13.5	84.9	15.5
Stereo4D only	11.3	88.0	8.00	91.8	5.52	95.8	7.50	92.3	5.65	94.8	6.77	93.7	9.29	91.4	16.0	73.4	15.0	82.2	15.6
AerialMD only	9.81	90.2	7.39	92.9	5.59	95.7	7.41	92.6	5.65	94.9	6.70	93.8	8.60	93.4	16.8	78.9	14.6	85.5	14.5
MegaSc. only	9.78	90.2	7.32	92.8	5.45	95.8	7.45	91.9	5.51	95.0	6.75	93.9	8.68	91.5	15.9	73.0	13.7	82.8	15.0
AerialMD.+MegaSc.	9.73	90.6	7.33	92.1	5.52	95.7	7.40	92.1	5.52	94.9	6.69	93.9	8.36	92.5	16.0	74.2	13.2	85.1	14.9
Evaluated on MetricScenes Test Set																			
All	5.24	97.0	3.67	97.9	-	-	4.02	97.1	2.98	98.2	4.15	97.1	7.59	93.4	26.5	73.8	26.0	79.1	-
Stereo4D only	6.23	96.6	3.83	97.6	-	-	4.37	97.0	3.04	98.1	4.17	97.0	10.6	88.2	34.0	39.3	33.5	45.6	-
AerialMD only	6.33	95.6	4.39	97.1	-	-	4.55	96.0	3.25	97.7	4.48	96.4	11.5	87.8	41.0	58.7	38.3	45.6	-
MegaSc. only	6.68	94.5	4.61	96.8	-	-	5.13	94.9	3.59	97.2	4.82	96.0	9.99	89.8	32.4	53.0	28.7	61.9	-
AerialMD.+MegaSc.	6.66	95.2	4.52	97.0	-	-	4.86	95.5	3.43	97.5	4.66	96.2	9.57	91.0	32.7	65.0	26.4	74.7	-

Fine-tuning on Different Subsets. The ablation results in Tab. 2 validate the multi-source design of the MetricScenes dataset by showing how each subset contributes to performance. Each ablated model is trained for 3 epochs, consistent with the WildMoGe training schedule. On the standard benchmarks, we observe that each data source addresses a specific aspect of the model’s performance. Specifically, Stereo4D helps improve boundary sharpness ($F_1 = 15.6$); AerialMegaDepth serves as a reliable metric anchor, resulting in strong metric-scale geometry accuracy; MegaScenes adds important perspective diversity, and when combined with AerialMegaDepth, achieves the lowest scale-invariant point error, which helps the model handle many different camera viewpoints. Generally, we find that joint training on the full corpus leads to better metric geometry and sharper boundaries. On the MetricScenes test set, we find that training exclusively on each subset tends to limit overall performance. In contrast, joint training across all subsets achieves superior results across every metric, suggesting a strong synergistic effect.

5 Conclusion

In this work, we identify scale collapse as a challenge in monocular metric depth/geometry estimation on in-the-wild scenes. Despite the impressive performance of modern models on standard benchmarks, they often fail to maintain consistent real-world dimensions on certain types of data, such as distant landmarks or vast landscapes. We consider this fundamentally a data problem rooted in the limited scale diversity of existing training sets, which are typically restricted by hardware constraints or a lack of semantic complexity. To bridge this gap, we propose MetricScenes, a diverse and metrically-grounded real-world dataset curated from widely available sources including Internet photo collections and online stereo videos/imagery. Through qualitative and quantitative evaluations, we found that fine-tuning the state-of-the-art MoGe-2 model on our dataset effectively overcomes scale-collapse, enabling faithful metric scale recovery in unconstrained environments while maintaining competitive performance on standard benchmarks.

6 Acknowledgment

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project), and the Major Program of the National Natural Science Foundation of China (Grant No. 62595772).

References

1. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *ArXiv abs/2111.08897* (2021)
2. Bhat, S.F., Birkel, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288* (2023)
3. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. *arXiv* (2024), <https://arxiv.org/abs/2410.02073>
4. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) *European Conf. on Computer Vision (ECCV)*. pp. 611–625. Part IV, LNCS 7577, Springer-Verlag (Oct 2012)
5. Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Doppelgangers: Learning to disambiguate images of similar structures. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 34–44 (2023)
6. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items (2022)
7. Duisterhof, B.P., Žust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: A fully-integrated solution for unconstrained structure-from-motion. *2025 International Conference on 3D Vision (3DV)* pp. 1–10 (2024), <https://api.semanticscholar.org/CorpusID:272988049>
8. Fonder, M., Droogenbroeck, M.V.: Mid-air: A multi-modal dataset for extremely low altitude drone flights. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* pp. 553–562 (2019), <https://api.semanticscholar.org/CorpusID:156052231>
9. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
10. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(12), 10579–10596 (2024)
11. Jin, L., Tucker, R., Li, Z., Fouhey, D., Snavely, N., Holynski, A.: Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. In: *CVPR* (2025)
12. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., 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*. pp. 780–791 (2023)

13. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kotschieder, P.: MapAnything: Universal feed-forward metric 3D reconstruction. In: International Conference on 3D Vision (3DV). IEEE (2026)
14. Koch, T., Liebel, L., Fraundorfer, F., Körner, M.: Evaluation of cnn-based single-image depth estimation methods. In: Leal-Taixé, L., Roth, S. (eds.) Proceedings of the European Conference on Computer Vision Workshops (ECCV-WS). pp. 331–348. Springer International Publishing (2019)
15. Koch, T., Liebel, L., Körner, M., Fraundorfer, F.: Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding (CVIU)* **191**, 102877 (2020). <https://doi.org/10.1016/j.cviu.2019.102877>
16. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision. pp. 71–91. Springer (2024)
17. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3182–3192 (2023), <https://api.semanticscholar.org/CorpusID:263135139>
18. Li, Y., Xiangli, Y., Averbuch-Elor, H., Snavely, N., Cai, R.: Long-tail internet photo reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2026)
19. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Computer Vision and Pattern Recognition (CVPR) (2018)
20. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. *ArXiv* **abs/2511.10647** (2025), <https://api.semanticscholar.org/CorpusID:282992334>
21. Mehl, L., Schmalfluss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
22. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: ECCV (2012)
23. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Gool, L.V.: Unidepthv2: Universal monocular metric depth estimation made simpler (2025), <https://arxiv.org/abs/2502.20110>
24. Roberts, M., Paczan, N.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. 2021 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10892–10902 (2020), <https://api.semanticscholar.org/CorpusID:226254406>
25. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
27. Schöps, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle adjusted direct RGB-D SLAM. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
28. Sun, P., Kretschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H.,

- Timofeev, A., Ettinger, S.M., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2443–2451 (2019), <https://api.semanticscholar.org/CorpusID:209140225>
29. Tung, J., Chou, G., Cai, R., Yang, G., Zhang, K., Wetzstein, G., Hariharan, B., Snavely, N.: Megascenes: Scene-level view synthesis at scale. *ArXiv abs/2406.11819* (2024)
 30. Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., Geiger, A.: Sparsity invariant cnns. In: International Conference on 3D Vision (3DV) (2017)
 31. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR abs/1908.00463* (2019), <http://arxiv.org/abs/1908.00463>
 32. Vuong, K., Ghosh, A., Ramanan, D., Narasimhan, S., Tulsiani, S.: Aerialmegadept: Learning aerial-ground reconstruction and view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
 33. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546* (2025), <https://arxiv.org/abs/2507.02546>
 34. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.A.: Tartanair: A dataset to push the limits of visual slam. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) pp. 4909–4916 (2020), <https://api.semanticscholar.org/CorpusID:214727835>
 35. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: Pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347* (2025)
 36. Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36–54. Springer (2024)
 37. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundation-stereo: Zero-shot stereo matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5249–5260 (2025)
 38. Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J.K., Ramanan, D., Hays, J.: Argoverse 2: Next generation datasets for self-driving perception and forecasting. *ArXiv abs/2301.00493* (2023)
 39. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. *ArXiv abs/1810.08705* (2018), <https://api.semanticscholar.org/CorpusID:53047282>
 40. Xiangli, Y., Cai, R., Chen, H., Byrne, J., Snavely, N.: Doppelgangers++: Improved visual disambiguation with geometric 3d features (2025)
 41. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024)
 42. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendmvs: A large-scale dataset for generalized multi-view stereo networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1787–1796 (2019), <https://api.semanticscholar.org/CorpusID:208248003>
 43. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. 2023 IEEE/CVF International Conference on Computer Vi-

- sion (ICCV) pp. 12–22 (2023), <https://api.semanticscholar.org/CorpusID:261064784>
44. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 9009–9019 (2023), <https://api.semanticscholar.org/CorpusID:259991083>
 45. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3d scene shape from a single image. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 204–213 (2020), <https://api.semanticscholar.org/CorpusID:229298063>