Argus: Metric Panoramic 3D Reconstruction for Indoor Scenes

Xi Li[1], Linyuan Li[1], Yan Wu[1], Tong Rao[1], Kai Zhang[1], Xinchen Hui[1], and Cihui Pan[1,2⋆]

[1] Realsee, China
[2] Quanzhou University of Information Engineering, China

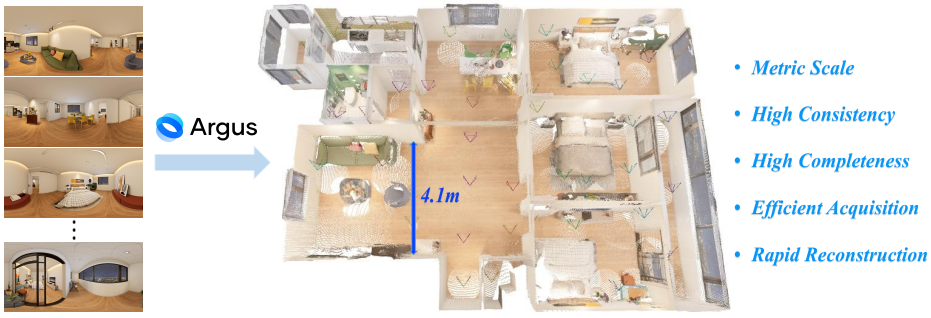

Fig. 1: Argus is a feed-forward 3D reconstruction network trained on our large-scale indoor panoramic 3D dataset Realsee3D. Given acquired sparse multi-view panoramic images, it rapidly reconstructs complete, consistent, and metric-scale 3D scenes.

Abstract. Metric feed-forward 3D reconstruction for panoramic data remains under-explored due to the lack of large-scale panoramic RGB-D training data. We present Realsee3D, a hybrid dataset of 10K indoor scenes (1K real, 9K synthetic) with 299K panoramic viewpoints and precise metric annotations, and Argus, a feed-forward network trained on it for metric panoramic 3D reconstruction. In the sparse unordered capture setting of Realsee3D, a poorly chosen coordinate anchor can cause global pose drift. Argus addresses this with a learned covisibility module that selects the geometrically optimal reference view to anchor the metric world frame. To further improve multi-task learning, we decompose the bidirectional pixel-to-world mapping into interpretable sub-steps with per-step supervision and cross-coordinate joint constraints, reinforcing geometric consistency across prediction branches. On the Realsee3D benchmark, Argus achieves state-of-the-art metric performance in camera pose estimation, depth estimation, and point cloud reconstruction. Project page: `https://argus-paper.realsee.ai`.

Keywords: Metric · Panoramic · Feed-Forward · 3D Reconstruction

⋆ Corresponding author: `pancihui001@realsee.com`

1 Introduction

Image-based 3D reconstruction is a fundamental problem in computer vision, widely applied to robotics [3,35,44], AR/VR [16,40], autonomous driving [26,59], visual localization [20,33,80], and other fields [49]. Recently, Transformer-based feed-forward 3D reconstruction has achieved breakthroughs in perspective image scenarios [52,84,89]. Extending such methods to panoramic imagery is particularly appealing: panoramic cameras are increasingly adopted for real-world indoor 3D capture [2], as a single panorama covers the full field of view and provides substantial cross-view overlap, aligning well with the sparse-view feed-forward paradigm by enabling rapid scene-level reconstruction from only a few captures. Despite its high practical value, metric feed-forward 3D reconstruction tailored for panoramic data remains largely unaddressed. Mainstream feed-forward models such as VGGT [84] possess multi-view geometric reasoning capabilities, but are trained on perspective RGB-D datasets and suffer severe performance degradation when directly applied to panoramic data. A key bottleneck lies in the scarcity of large-scale, high-quality panoramic 3D training data with accurate metric annotations. To fill this gap, this paper presents Realsee3D, a large-scale indoor multi-view panoramic RGB-D dataset with room-level coverage and precise metric annotations, and Argus, a feed-forward network trained on Realsee3D that integrates learnable covisibility-guided reference selection with geometric factorization supervision for metric panoramic 3D reconstruction.

The Realsee3D dataset aligns with real-world panoramic capture scenarios, where images are sparsely sampled and provided as an unordered set. In this practical setting, existing feed-forward reconstruction methods [45, 84] anchor the global coordinate system to a pre-defined heuristic reference view. When the anchor is an isolated or boundary viewpoint, insufficient cross-view covisibility degrades geometric constraints, causing severe global pose drift and spatially inconsistent reconstruction. π^3 [91] adopts permutation-equivariant architectures to address permutation sensitivity, but the tightly coupled $N \times N$ pairwise constraints require a fixed-reference branch to stabilize training from scratch. Traditional methods [71] typically rely on hand-designed heuristics to select a robust world coordinate reference, e.g., the highest-feature view or the pair with most matches. Inspired by these methods, we propose a learnable covisibility module as an alternative. Instead of a fixed heuristic anchor, our module predicts global inter-view connectivity to dynamically select the optimal reference frame, suppressing drift from degenerate viewpoints and improving cross-view consistency. The pipeline retains approximate permutation equivariance, while inheriting the smooth optimization landscape of reference-frame-oriented metric learning.

Beyond reference selection, effective multi-task supervision is critical for reconstruction quality. Recent works [84, 87–89] show that joint supervision over point maps, poses, and depth yields strong geometric synergies. Building on this insight, we propose an overcomplete geometric factorization supervision strategy that decomposes pixel-to-world transformations under the panoramic model into interpretable sub-steps, each supervised individually and jointly across coordinate frames. This lowers optimization difficulty while enhancing multi-task

synergy. We further observe empirically that sufficient high-quality metric data, combined with expressive modeling, suppresses ERP distortion and boundary artifacts without task-specific designs.

In summary, the main contributions of this work include:

- A large-scale indoor 3D dataset Realsee3D, and a data-driven feed-forward metric panoramic 3D reconstruction network Argus.
- A covisibility-based reference view learning method that anchors the metric coordinate system and enhances reconstruction robustness to pose drift.
- An overcomplete geometric factorization supervision strategy that decomposes pixel-to-world transformations into supervised sub-steps with cross-coordinate consistency constraints, boosting multi-task synergy.
- A new panoramic 3D reconstruction benchmark on Realsee3D, on which Argus achieves best overall performance across all evaluated tasks.

2 Related Work

2.1 Traditional 3D Reconstruction

Traditional 3D reconstruction relies on optimization-based pipelines with explicit geometric modeling. Classical Structure-from-Motion (SfM) pipelines [1, 61, 71] recover camera poses and sparse point clouds through a multi-stage process: inter-view feature matching [58, 69], robust two-view geometry estimation with RANSAC [6] and other estimators [8, 78], view graph construction [10, 62], and global bundle adjustment refinement [32, 81]. Given the recovered poses, Multi-View Stereo (MVS) densifies geometry via multi-view photometric reasoning, spanning classical pipelines [28, 29, 72, 74] and learning-based approaches [36, 92, 98]. Recent efforts increasingly integrate learned components into the SfM/MVS pipeline, including keypoint detectors [21, 34, 102], learned matchers [54, 70], detector-free semi-dense [50,77,90] and dense matching [24,25], monocular depth priors [52,87,96], and end-to-end frameworks [23,37,95]. While these hybrid approaches achieve high accuracy, they still inherit multi-stage iterative optimization [41, 86, 101], incurring prohibitive cost for large image sets.

2.2 Feed-Forward 3D Reconstruction

Feed-forward methods [13, 56, 84] eliminate iterative optimization and directly predict scene geometry and camera motion from input images. Transformer architectures [17, 18, 83] have emerged as the dominant backbone for this task. DUSt3R [89] redefines feed-forward geometry prediction as dense point map regression and achieves robust pairwise reconstruction. VGGT [84] introduces alternating attention and multi-task supervision for multi-view geometry learning. MapAnything [45] expands this paradigm to universal metric reconstruction with support for multi-modal prior inputs. However, most multi-view feed-forward methods [84,100,107] remain highly sensitive to input view order due to their implicit coordinate anchoring to a selected reference view. π^3 [91] eases this

sensitivity via a permutation-equivariant design for unordered input robustness. More recently, VGGT-Ω [85] presents the first empirical validation of the scaling law for feed-forward 3D reconstruction. Despite these rapid advances, existing feed-forward methods are developed and evaluated exclusively on perspective imagery, leaving panoramic scenes largely unexplored. Our work bridges this gap by introducing a feed-forward model for metric panoramic reconstruction.

2.3 Panoramic 3D Reconstruction and Datasets

The rapid progress of feed-forward methods [27, 75, 84] underscores that high-quality 3D datasets [9, 55, 67, 99] are the cornerstone of robust feed-forward 3D reconstruction [15, 19, 51]. However, most large-scale indoor RGB-D datasets [4, 76, 79, 105] are built on perspective projection [30, 64, 68, 93], leaving a critical gap in dedicated training data and standardized benchmarks for panoramic metric reconstruction. Existing panoramic 3D datasets such as Matterport3D [12] and Stanford2D3D [5] primarily serve as evaluation benchmarks for tasks like monocular depth estimation [31, 53, 65] and semantic segmentation [106], but their limited scale and annotation schemes cannot support learning-based multi-view metric reconstruction. Consequently, panoramic monocular depth estimation methods [11, 38, 48] typically resort to transfer learning from foundation models pre-trained on large-scale perspective images such as DepthAnything V2 [97]. To advance this under-explored field, we introduce Realsee3D, a large-scale multi-view panoramic RGB-D dataset with precise metric annotations, enabling training and evaluation of feed-forward panoramic reconstruction models and flexible extension to downstream tasks including depth completion [94], novel view synthesis [39, 46, 60], and embodied intelligence [104], among others.

3 Realsee3D Dataset

3.1 Dataset Overview and Construction

We introduce Realsee3D[3], a large-scale high-resolution panoramic dataset with metric annotations combining photorealistic real-world captures and diverse synthetic scenes, containing 10,000 indoor scenes, 95,962 rooms, 299,073 panoramic viewpoints, and two complementary subsets detailed below.

Real Subset. This subset consists of 1,000 real-world residential scenes, covering 9,483 rooms and 24,263 panoramic RGB-D viewpoints. We acquire the data using tripod-mounted Realsee Galois 3D LiDAR cameras with nearly co-centered RGB and LiDAR sensors, which greatly simplifies image stitching and ensures geometric consistency between appearance and depth measurements. At each capture position, the camera rotates to five preset orientations. To preserve photometric fidelity, we stitch images directly in the RAW domain and adopt HDR multi-exposure stacking to handle complex indoor lighting. After on-site initial pose estimation, we perform global registration and pose refinement, which

[3] Dataset available at `https://dataset.realsee.ai`.

Table 1: Comparison with Existing Indoor Panoramic Datasets. We distinguish unique viewpoints (spatial capture locations) from total images for accurate comparison. Realsee3D offers an unprecedented scale.

Dataset	Type	Scenes	Rooms	Unique Viewpoints	Images	Modalities
Stanford2D3D [5]	Real	6	270	1,413	1,413	RGB-D, Pose
Matterport3D [12]	Real	90	2,056	10,800	10,800	RGB-D, Pose
ZInD [14]	Real (Unfurnished)	1,524	–	71,474	71,474	RGB, Pose
Structured3D [103]	Synthetic	3,500	21,835	21,835	196,515	RGB-D, Pose
Realsee3D (Real)	Real (Furnished)	1,000	9,483	24,263	24,263	RGB-D, Pose
Realsee3D (Synthetic)	Synthetic	9,000	86,479	274,810	274,810	RGB-D, Pose
Realsee3D (Total)	**Hybrid**	**10,000**	**95,962**	**299,073**	**299,073**	**RGB-D, Pose**

jointly optimizes visual and 3D features across all scans via covisibility, with manual corrections for accurate trajectory alignment when necessary. The final output includes high-resolution ERP RGB images, sparse metric depth maps, and precise camera poses, faithfully capturing real-world illumination, textures, and object distributions.

Synthetic Subset. To further scale the dataset and enrich scene diversity, we construct a synthetic subset of 9,000 procedurally generated scenes, covering 86,479 rooms and 274,810 panoramic viewpoints. All scenes are generated through a multi-stage pipeline grounded in real-world floorplan distributions. We adopt a hybrid rule- and learning-based discrete optimization framework that integrates real-scene layout priors and expert design knowledge to produce plausible scene configurations. We leverage a proprietary 3D asset library containing over 100,000 high-fidelity models across 200+ semantic categories. These models preserve fine-grained geometry and use physically-based rendering (PBR) materials, effectively reducing the synthetic-real domain gap. After layout generation, a graph-based viewpoint selection algorithm determines optimal capture positions by simulating real-world roaming trajectories, ensuring full and high-quality coverage of the 3D environment and yielding dense multi-view observations per room. Finally, we render scenes at these viewpoints using a customized Unreal Engine 5 pipeline with hardware-accelerated ray tracing, which simulates complex light transport including multi-bounce global illumination and ambient occlusion. Each scene is rendered under five distinct illumination schemes (Warm Day, Cold Day, Natural Day, Warm Night, Cold Night) to maximize intra-scene diversity. This pipeline produces dense, high-fidelity RGB-D data and precise camera poses, providing perfect geometric ground-truth free from sensor noise in real scans. The hybrid design of Realsee3D ensures both photorealistic diversity and accurate geometric annotations, making it substantially larger and more diverse than existing panoramic datasets for dense 3D reconstruction.

3.2 Comparison with Existing Datasets

As shown in Table 1, we compare Realsee3D against several representative indoor panoramic datasets. While Stanford2D3D [5] and Matterport3D [12]

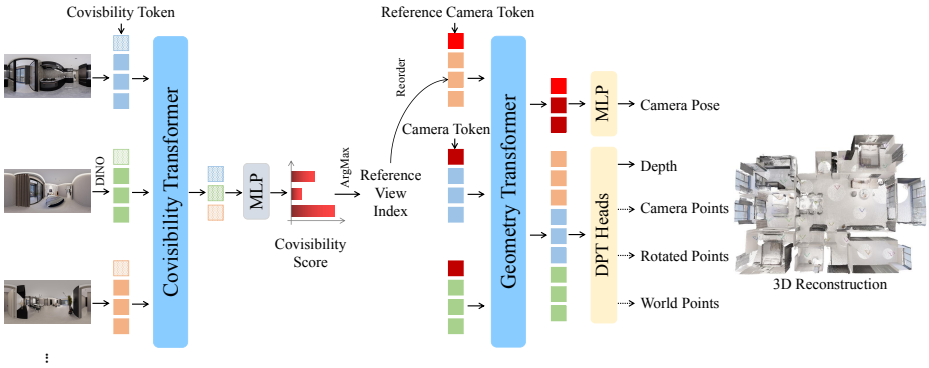

Fig. 2: Overview of Argus. We first select the optimal reference frame via a Covisibility Transformer, then aggregate multi-view geometric features through a reference-based Geometry Transformer, followed by multiple prediction branches for intermediate geometric decomposition representations that facilitate multi-task learning.

lay the foundation for indoor scene understanding with high-quality real-world panoramic data, they are limited in scene count. ZInD [14] provides massive real panoramas, yet focuses on unfurnished scenes and lacks dense metric depth. Structured3D [103] presents large-scale synthetic data with fine annotations, yet its 196,515 images are captured from only 21,835 unique spatial viewpoints (one per room) under varying lighting and furniture configurations. Realsee3D complements these efforts as a large-scale hybrid dataset with dense spatial coverage, establishing a robust new benchmark for metric 3D reconstruction.

4 Method

4.1 Overview of Argus

An overview of our network is shown in Fig. 2. Given unordered panoramas, our model first leverages DINOv2 [63] to generate patch tokens. Covisibility tokens are added to per-view tokens and fed into a lightweight Covisibility Transformer with $L_c = 2$ alternating attention layers [84] (self-attention within each view alternated with self-attention across views) and then fed into an MLP layer to predict covisibility scores. The highest-scoring view is chosen as the reference frame with a dedicated reference camera token, while other views use standard camera tokens. The aggregated token sequence is then processed by a Geometry Transformer with $L_g = 24$ alternating attention layers. Finally, camera poses are regressed from camera tokens by an MLP layer that outputs a 9-dimensional vector per view: a 3D translation, a 4D quaternion rotation, and 2 confidence scores (one for rotation, one for translation), and depth and point maps across coordinate systems are predicted from image tokens via distinct DPT [66] heads. Each DPT head outputs an extra channel for pixel-wise confidence. The depth

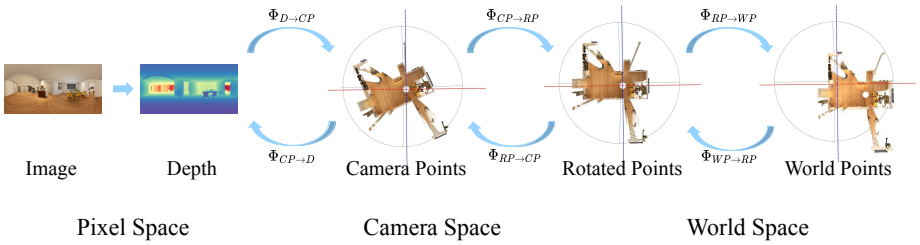

Fig. 3: Pixel-to-World Bidirectional Transformations for a Single View.

head is activated with $\exp(\cdot)$ to ensure positivity, point map heads use an inverse log transform $f(x) = \text{sign}(x) \cdot (\exp(|x|) - 1)$ to handle the wide dynamic range of 3D coordinates, where x denotes the raw head output, and all confidence values are activated via $1 + \exp(\cdot)$ to guarantee a lower bound of 1.

4.2 Learning to Select Reference View

To enable the model to learn optimal reference frames, we construct the supervision labels as follows. Let the image set be $\mathbb{I} = \{I_1, I_2, \ldots, I_N\}$, where N is the total number of images. We precompute the covisibility scores of all image pairs based on pose and depth and then construct a global covisibility matrix $\mathbf{M}$. Afterwards, Dijkstra's algorithm [22] is adopted to select the optimal reference view $\mathcal{I}$. More details are provided in the supplementary materials. The loss is computed by measuring the discrepancy between predicted covisibility logits $\hat{\mathbf{C}} \in \mathbb{R}^N$ and the one-hot ground-truth $\mathbf{C} \in \mathbb{R}^N$. The ground-truth sets $\mathbf{C}_{\mathcal{I}} = 1$ for the optimal view $\mathcal{I}$ with maximum global covisibility and 0 otherwise ($\mathcal{I} \in \{1, \ldots, N\}$). We adopt binary cross-entropy (BCE) loss for supervision during training. During inference, we select the reference frame $\hat{\mathcal{I}}$ with the highest score via the arg max operation, which provides approximate permutation equivariance since argmax is independent of input ordering. For implementation simplicity, we swap the tokens of this view with those of a fixed reference frame (the first frame) to achieve an equivalent effect.

4.3 Panoramic Geometric Factorization

Directly regressing world-coordinate point maps from images conflates multiple geometric transformations into a single prediction step, which increases optimization difficulty and limits cross-task supervisory signals. To address this, we predict each intermediate representation independently using distinct DPT [66] heads and supervise intermediate representations of forward and inverse pixel-to-world coordinate transformations in panoramic projections via both independent and joint cross-coordinate geometric constraints (refer to Sec. 4.5).

For $W \times H$ ERP panoramas, 2D pixel coordinates (u, v) map directly to spherical latitude $\theta = \left(\frac{v}{H} - 0.5\right) \cdot \pi$ and longitude $\phi = \left(\frac{u}{W} - 0.5\right) \cdot 2\pi$, enabling

computation of 3D point clouds on the unit sphere as:

$$P_u = \begin{bmatrix} \cos(\theta) \cdot \sin(\phi) \\ \sin(\theta) \\ \cos(\theta) \cdot \cos(\phi) \end{bmatrix} \tag{1}$$

All forward and inverse pixel-to-world transformations Φ for a single view are shown in Fig. 3, which can be decomposed into the following:

$$\begin{cases} \Phi_{D\to CP} : P_c = D \odot P_u, \\ \Phi_{CP\to RP} : P_r = RP_c, \\ \Phi_{RP\to WP} : P_w = P_r + \boldsymbol{t}, \\ \Phi_{WP\to RP} : P_r = P_w - \boldsymbol{t}, \\ \Phi_{RP\to CP} : P_c = R^{-1}P_r, \\ \Phi_{CP\to D} : D = \|P_c\|_2 \end{cases} \tag{2}$$

Here D is the depth map of panoramic image I, R and $\boldsymbol{t}$ are the rotation matrix and translation vector of the camera pose, respectively. In the forward direction, the camera point map $P_c = D \odot P_u$ scales the unit sphere by depth, $P_r = RP_c$ rotates into the reference frame, and $P_w = P_r + \boldsymbol{t}$ translates to world coordinates; the inverse direction reverses these steps. In $\Phi_{CP\to D}$, the angular reprojection error is neglected for efficiency, as its influence on results is negligible.

4.4 Metric Prediction

Unlike perspective cameras where focal length and depth are inherently coupled, the fixed ERP projection model eliminates intrinsic ambiguity and thus simplifies the model's learning of metric depth. Combined with the consistent metric scale across our dataset (from LiDAR capture and physically based rendering), this enables direct metric prediction without a separate scale estimation module [45, 88]. We simply normalize the ground-truth scale (e.g., divide by $s = 10$) during supervision, which yields stable and accurate metric predictions.

4.5 Loss Function

We train Argus end-to-end using multiple losses, including a covisibility loss, camera loss, depth loss, multiple point map losses, and a geometry joint loss.
Covisibility Loss. The covisibility loss $\mathcal{L}_{\text{covis}}$ is formulated as:

$$\mathcal{L}_{\text{covis}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\mathbf{C}_i \log(\sigma(\hat{\mathbf{C}}_i)) + (1-\mathbf{C}_i)\log(1-\sigma(\hat{\mathbf{C}}_i))\Big] \tag{3}$$

where N is the sequence length and $\sigma(\cdot)$ is the sigmoid function.
Camera Loss. The camera loss employs two terms: the quaternion-based rotation loss $\mathcal{L}_q$ and the translation loss $\mathcal{L}_t$. These measure the L1 distance between

predicted and ground-truth quaternions $\hat{q}_i$, q_i, and between scaled predicted and scaled ground-truth translation vectors $\hat{t}_i$, $\frac{t_i}{s}$. Unlike previous methods [45, 84], additional confidence scores $\mathcal{C}_i^q$ and $\mathcal{C}_i^t$ are predicted respectively for rotation and translation to improve usability. These scores reflect the model's pose estimation confidence and are integrated into the corresponding loss terms as follows:

$$\mathcal{L}_q = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{C}_i^q + 1)\cdot\|\hat{q}_i - q_i\|_1 - \alpha\log\mathcal{C}_i^q \tag{4}$$

$$\mathcal{L}_t = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{C}_i^t + 1)\cdot\|\hat{t}_i - \frac{t_i}{s}\|_1 - \alpha\log\mathcal{C}_i^t \tag{5}$$

The overall camera loss is formulated as: $\mathcal{L}_{cam} = \mathcal{L}_q + \mathcal{L}_t$.

Depth Loss. The depth loss follows [84], adopting an aleatoric uncertainty loss that weights the L2 between the predicted depth $\hat{D}_{i,j}$ and the ground-truth depth $D_{i,j}$ with the predicted uncertainty map $C_{i,j}^D$. The $\mathcal{L}_d$ is formulated as:

$$\mathcal{L}_d = \frac{1}{NHW}\sum_{i=1}^{N}\sum_{j=1}^{HW}(C_{i,j}^D + 1)\left[\left\|\hat{D}_{i,j} - \frac{D_{i,j}}{s}\right\| + \left\|\nabla\hat{D}_{i,j} - \nabla\frac{D_{i,j}}{s}\right\|\right] - \alpha\log C_{i,j}^D \tag{6}$$

where ∇ is the gradient operator.

Point Map Loss. The point map loss $\mathcal{L}_p$ is analogous to the depth loss.

$$\begin{aligned}\mathcal{L}_p = \frac{1}{3NHW}\sum_{i=1}^{N}\sum_{j=1}^{HW}(C_{i,j}^P + 1)\Big[&\left\|\hat{P}_{i,j} - \frac{P_{i,j}}{s}\right\| \\ &+ \left\|\mathbf{n}(\hat{P}_{i,j}) - \mathbf{n}\left(\frac{P_{i,j}}{s}\right)\right\|\Big] - \alpha\log C_{i,j}^P\end{aligned} \tag{7}$$

Note that $\mathbf{n}(\cdot)$ denotes the operator computing point normals. Substituting predicted point maps under distinct spatial coordinate systems into Eq. (7) yields $\mathcal{L}_{cp} = \mathcal{L}_p(P_c)$, $\mathcal{L}_{rp} = \mathcal{L}_p(P_r)$ and $\mathcal{L}_{wp} = \mathcal{L}_p(P_w)$, respectively.

Geometry Joint Loss. To promote mutual enhancement across geometric prediction branches, we further supervise adjacent transformations in all forward and inverse steps of pixel-to-world coordinate mapping. This is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{joint}} = &\mathcal{L}_{\text{p}}\big[\Phi_{D\to CP}(\hat{D})\big] + \mathcal{L}_{\text{p}}\big[\Phi_{CP\to RP}(\hat{P}_c, \hat{R})\big] + \mathcal{L}_{\text{p}}\big[\Phi_{RP\to WP}(\hat{P}_r, \hat{t})\big] \\ &+ \mathcal{L}_{\text{p}}\big[\Phi_{WP\to RP}(\hat{P}_w, \hat{t})\big] + \mathcal{L}_{\text{p}}\big[\Phi_{RP\to CP}(\hat{P}_r, \hat{R})\big] + \mathcal{L}_{\text{d}}\big[\Phi_{CP\to D}(\hat{P}_c)\big]\end{aligned} \tag{8}$$

Each term takes the prediction from one branch, applies the predicted geometric transform to derive the next-stage representation, and supervises the result against the corresponding ground-truth, thereby enforcing cross-branch geometric consistency. Note that all inputs to the joint loss ($\hat{D}$, $\hat{P}_c$, $\hat{P}_r$, $\hat{P}_w$, $\hat{R}$, $\hat{\boldsymbol{t}}$) are network predictions from their respective branches, ensuring gradient flow across all branches for mutual facilitation.

Total Loss. The final weighted loss $\mathcal{L}$ is as follows:

$$\mathcal{L} = 0.1\cdot\mathcal{L}_{covis} + 5.0\cdot\mathcal{L}_{cam} + \mathcal{L}_d + \mathcal{L}_{cp} + \mathcal{L}_{rp} + \mathcal{L}_{wp} + \mathcal{L}_{joint} \tag{9}$$

5 Implementation Details

We train Argus by optimizing the training loss with the AdamW [57] optimizer for 99K iterations. We use a cosine learning rate scheduler with a peak learning rate of $5e-5$ and a warmup of 9.9K iterations. For every batch, we randomly sample 2 to 28 views from a random training scene. Since the covisibility adjacency matrix between views in each scene is precomputed, we ensure the sampled views are connected during training. Besides, we randomly split both real and synthetic subsets into training and test sets at a 9:1 scene-level ratio. Due to the non-uniform pixel distribution of ERP panoramic images, pixel utilization at the north and south poles is extremely low. We first resize the panoramic images, depths and point maps to 560×280, then crop the top and bottom 15% of pixels each, yielding a final resolution of 560×196. We also randomly apply color jittering, Gaussian blur, and grayscale augmentation to the views. In addition, a random rotation around the Y-axis is applied to the input of each panoramic viewpoint, which corresponds to a horizontal pixel shift in the ERP image and augments viewpoint yaw diversity. We set $\alpha = 0.2$ as the weight of the regularization term for all confidence losses. The training runs on 24 H20 GPUs with 141 GB memory over 36 hours. We employ gradient norm clipping with a threshold of 1.0 to ensure training stability. We leverage bfloat16 precision and gradient checkpointing to improve GPU memory and computational efficiency. All evaluations are completed on a single H20 GPU.

6 Benchmarking & Results

6.1 Baselines

We benchmark Argus on a comprehensive suite of 3D vision geometry tasks using our Realsee3D dataset. Given the long-standing scarcity of large-scale 3D panoramic datasets, no off-the-shelf feed-forward network is available for direct panoramic 3D reconstruction. For a fair comparison with SOTA methods, we therefore finetune VGGT [84] (without tracking supervision [42, 43]), MapAnything [45] (V1.1.1, using only image inputs) and π^3 [91] as our main baselines on Realsee3D. All models share consistent experimental settings, including the same train/test split, data preprocessing, and the same training iterations. We denote these adapted models as VGGT360, MapAnything360 and $\pi^3$360. Argus and VGGT360 are initialized from VGGT [84] pre-trained weights, others from their own checkpoints. It should be noted that the released π^3 pre-training weights are initialized from VGGT. Realsee3D contains scenes represented by unordered images of arbitrary quantity. For comprehensive scene evaluation, validation uses all per-scene panoramic images, differing from prior work [84] sampling fixed frames from ordered video streams. Moreover, due to the challenging dataset, traditional open-source SfM methods (COLMAP [71] and OpenMVG [61]), when applied out of the box, are too slow and yield unusable results, so they are omitted from comparison.

Table 2: Camera Pose Estimation on the Realsee3D Dataset.

Dataset Part	Method	AUC↑			RRA↑			RTA↑			ATE↓	A.R.↑
		@5°	@10°	@20°	@5°	@10°	@20°	@5°	@10°	@20°		
Real	VGGT360 [84]	67.89	83.00	91.00	96.88	99.10	99.46	97.37	99.05	99.66	–	–
	MapAnything360 [45]	**72.66**	85.37	92.45	**99.41**	99.94	**100.00**	95.86	98.97	99.77	0.134	94.0
	$\pi^3$360 [91]	72.39	**85.92**	**92.90**	98.60	**100.00**	**100.00**	**98.58**	**99.74**	**99.92**	–	–
	Ours	71.88	85.52	92.64	98.30	**100.00**	**100.00**	98.10	99.55	99.83	**0.096**	**98.2**
Synthetic	VGGT360 [84]	91.10	95.38	97.60	99.67	99.86	99.92	99.57	99.80	99.89	–	–
	MapAnything360 [45]	90.84	95.11	97.42	99.46	99.78	99.92	99.20	99.68	99.83	0.087	99.3
	$\pi^3$360 [91]	93.46	96.58	98.20	99.72	99.88	99.91	99.62	**99.82**	99.89	–	–
	Ours	**94.44**	**97.29**	**98.63**	**99.76**	**99.90**	**99.95**	**99.65**	**99.82**	**99.90**	**0.027**	**99.8**

Table 3: Multi-view Depth Estimation on the Realsee3D Dataset.

Alignment	Method	Real					Synthetic				
		AbsRel↓	δ_1↑	δ_2↑	RMSE↓	MAE↓	AbsRel↓	δ_1↑	δ_2↑	RMSE↓	MAE↓
IRLS	VGGT360 [84]	0.051	64.22	96.21	0.445	0.119	0.025	86.60	98.78	0.158	0.035
	MapAnything360 [45]	0.063	54.75	95.72	0.541	0.126	0.034	76.47	98.65	0.165	0.042
	$\pi^3$360 [91]	0.053	58.63	95.87	0.454	0.129	0.033	78.14	98.25	0.192	0.052
	Ours	**0.048**	**70.22**	**96.51**	**0.401**	**0.102**	**0.019**	**91.42**	**99.03**	**0.138**	**0.027**
Median	VGGT360 [84]	0.056	57.69	96.08	0.447	0.127	0.029	85.27	98.76	0.159	0.038
	MapAnything360 [45]	0.066	51.08	95.90	0.563	0.132	0.037	74.62	98.66	0.186	0.046
	$\pi^3$360 [91]	0.060	51.85	95.65	0.456	0.140	0.034	75.96	98.22	0.194	0.055
	Ours	**0.050**	**65.81**	**96.60**	**0.406**	**0.107**	**0.019**	**90.75**	**99.02**	**0.139**	**0.029**
Metric	MapAnything360 [45]	0.070	46.31	95.79	0.567	0.137	0.063	74.64	98.64	0.185	0.044
	Ours	**0.050**	**66.28**	**96.62**	**0.407**	**0.106**	**0.035**	**91.24**	**99.02**	**0.139**	**0.027**

6.2 Camera Pose Estimation

We evaluate camera pose prediction on both subsets of Realsee3D, reporting AUC, Relative Rotation Accuracy (RRA), and Relative Translation Accuracy (RTA) at thresholds of 5°, 10°, 20°, Absolute Trajectory Error RMSE (ATE) for global pose accuracy, and Acceptance Rate (A.R.)—the proportion of scenes where all poses have rotation error $< 10°$ and translation error $< 0.5\,\mathrm{m}$—to assess practical usability. As shown in Tab. 2, Argus achieves the best metric pose accuracy (ATE and A.R.) on both subsets and leads in AUC on the synthetic subset. Although $\pi^3$360 shows marginally higher relative metrics on the real subset—likely due to its stronger pre-training prior—Argus leads decisively in global metric accuracy and dominates all metrics on the synthetic subset where the domain gap from pre-training data is smaller. Compared with MapAnything360 (which supports metric prediction), Argus reduces ATE by 28% on the real subset and 69% on the synthetic subset.

6.3 Depth Estimation

We benchmark multi-view depth estimation on the Realsee3D dataset using Absolute Relative Error (AbsRel), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and interior percentage metrics $\delta_{<1.03}$ (δ_1) and $\delta_{<1.25}$ (δ_2). As several baselines cannot predict metric depth, ground-truth alignment is necessary before evaluation. For fair comparison with all methods, we report results under three alignment schemes: Iterative Reweighted Least Squares

Table 4: Point Map Reconstruction on the Realsee3D Dataset.

Alignment	Method	Real						Synthetic					
		Acc.↓		Comp.↓		N.C.↑		Acc.↓		Comp.↓		N.C.↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
ICP	VGGT360 [84]	**0.058**	0.039	0.074	**0.034**	0.884	0.981	0.023	0.014	0.016	0.011	0.923	0.995
	MapAnything360 [45]	0.082	0.054	0.127	0.095	0.829	0.937	0.038	0.026	0.053	0.043	0.880	0.982
	$\pi^3$360 [91]	**0.058**	0.040	0.078	**0.034**	0.868	0.975	0.030	0.017	0.018	0.013	0.888	0.991
	Ours	**0.058**	**0.037**	**0.062**	**0.034**	**0.894**	**0.983**	**0.018**	**0.011**	**0.014**	**0.010**	**0.943**	**0.997**
Metric	MapAnything360 [45]	0.089	0.057	0.075	0.047	0.825	0.962	0.039	0.023	0.025	0.016	0.873	0.991
	Ours	**0.056**	**0.037**	**0.063**	**0.034**	**0.894**	**0.986**	**0.020**	**0.012**	**0.015**	**0.010**	**0.939**	**0.997**

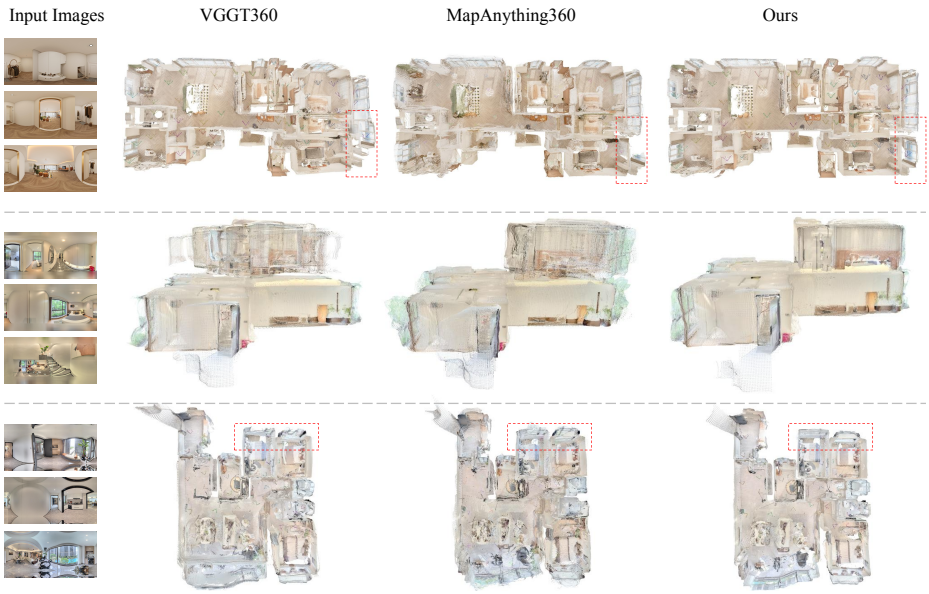

Fig. 4: Qualitative Comparison.

(IRLS) alignment [47], median alignment, and absolute metric evaluation without any alignment. Results are given in Tab. 3. Argus outperforms all baselines in all metrics. We further evaluate Argus on monocular depth estimation for Realsee3D, and test its zero-shot generalization on two standard benchmarks: Matterport3D [12] and Stanford2D3D [5]. Although not specifically trained for monocular depth estimation, our method still achieves highly competitive performance. Please refer to supplementary materials for the detailed results.

6.4 Point Map Reconstruction

We evaluate scene-level point map reconstruction in the world coordinate system on our Realsee3D dataset, reporting Accuracy (Acc.), Completeness (Comp.), and Normal Consistency (N.C.) in Tab. 4. We present results under two alignment schemes: alignment using the Umeyama algorithm [82] followed by Iterative

Table 5: Runtime and Peak GPU Memory Usage of Argus with Varying Input Frame Counts.

# Image Number	1	2	4	8	16	32	64	128
Runtime (s)	0.11	0.13	0.18	0.32	0.66	1.47	3.67	10.55
Memory (GB)	2.05	2.24	2.60	3.38	3.54	3.88	4.55	7.22

Closest Point (ICP) refinement [7], and absolute metric evaluation without any alignment. Argus achieves overall superior alignment performance and holds a clear absolute advantage under metric evaluation. The qualitative comparison of reconstructions is shown in Fig. 4, Argus yields more accurate metric geometry and sharper structural boundaries than previous methods.

6.5 Efficiency Evaluation

The runtime and peak GPU memory usage across different numbers of input frames are reported in Tab. 5. In practice, we run inference in BF16 precision and perform on-the-fly cleanup of redundant intermediate features, which keeps the peak memory footprint at a favorable level even when the number of input images exceeds 100. Note that the reported data ignore the memory of loading the pretrained model (4.64GB). Our model contains 1.31 billion parameters. Notably, the multiple DPT heads for cross-coordinate point map prediction can also be disabled during inference, as optimal performance is typically achieved with only the pose and depth heads. Thus, our multi-head prediction and supervision scheme markedly boosts geometric learning during training, with zero additional inference overhead.

6.6 Visualization

Reference View Learning. As shown in Fig. 5, real-world capture often starts from arbitrary positions, and the initial viewpoint is frequently located at scene corners or boundaries with low covisibility, which greatly raises the difficulty of robust reconstruction. With global covisibility supervision, Argus selects a superior initial reference view, typically near the scene center with stronger direct or indirect connections to other viewpoints. This improves scene-level reconstruction robustness by reducing drift and accumulated errors. More importantly, it enables metric prediction in the absolute world coordinate frame.

Generalizability. As shown in Fig. 6, Argus generalizes strongly beyond the training distribution. In particular, it maintains reliable metric 3D reconstruction for input panoramas with diverse AI-generated artistic styles [73] (e.g., sketches, cartoons). Furthermore, Argus exhibits strong robustness under challenging capture settings: weak covisibility, multi-floor indoor scenes, long-range viewpoints, and small-scale outdoor environments.

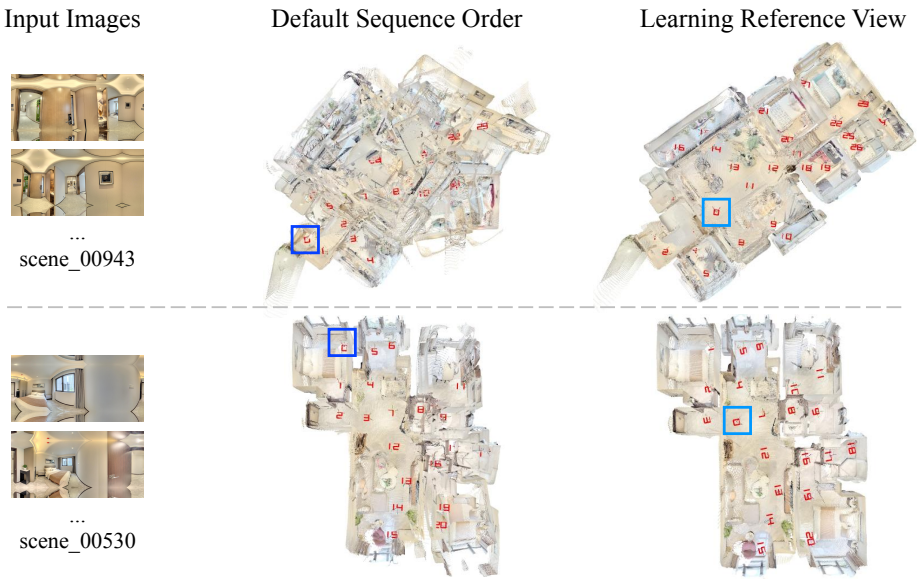

Fig. 5: Visualization of the Effectiveness of Reference View Learning.

Table 6: Ablation Studies on the Synthetic Subset of Realsee3D Dataset.

Method	AbsRel↓	δ_1↑	δ_2↑	RMSE↓	MAE↓	Acc.↓	Comp.↓	N.C.↑	ATE↓
Full	**0.035**	**91.24**	**99.02**	**0.139**	**0.027**	**0.020**	**0.015**	**0.939**	**0.027**
w/o reference view learning	0.037	90.45	98.96	0.142	0.029	0.024	0.017	0.923	0.060
w/o $\mathcal{L}_{joint}$	0.047	90.71	98.86	0.158	0.030	**0.020**	**0.015**	0.937	0.049
w/o $\mathcal{L}_{cp}$	0.055	89.90	98.79	0.161	0.031	**0.020**	**0.015**	0.936	0.047
w/o $\mathcal{L}_{rp}$	0.045	90.55	98.83	0.159	0.030	**0.020**	**0.015**	0.937	0.051
w/o $\mathcal{L}_{cp}$ & $\mathcal{L}_{rp}$	0.055	89.75	98.79	0.160	0.031	**0.020**	**0.015**	0.936	0.048
w/o $\mathcal{L}_{cp}$ & $\mathcal{L}_{rp}$ & $\mathcal{L}_{wp}$	0.075	88.52	98.60	0.169	0.034	–	–	–	0.061

6.7 Ablation Studies

Ablation studies are presented in Tab. 6. We report metric prediction results on the synthetic subset with key designs progressively ablated. Removing reference view learning more than doubles the ATE (0.027→0.060), confirming its critical role in establishing a consistent global coordinate frame. The geometric joint loss and multi-coordinate point map supervision improve both depth and pose accuracy: removing $\mathcal{L}_{joint}$ raises AbsRel from 0.035 to 0.047 and ATE from 0.027 to 0.049, and progressively dropping point map losses further degrades performance, with the removal of all three causing the largest decline (AbsRel 0.075, RMSE 0.169). Notably, removing $\mathcal{L}_{cp}$ alone has a larger impact than removing $\mathcal{L}_{rp}$ alone, suggesting that camera-coordinate supervision provides a stronger learning signal for depth. Although P_c and D are deterministically interconvertible, they represent geometry in different dimensional spaces and induce significantly different gradient dynamics during optimization.

Fig. 6: Visual Samples with Generalization Ability.

7 Conclusions, Limitations, and Future Work

We present Argus, a feed-forward model that achieves metric panoramic 3D reconstruction from unordered indoor images in a single forward pass, together with Realsee3D, a hybrid indoor benchmark of 10K real and synthetic scenes with 299K panoramic viewpoints and precise metric annotations. Argus introduces covisibility-guided reference view selection to robustly anchor unordered inputs, and a panoramic geometric factorization scheme that decomposes pixel-to-world mapping into independently supervised intermediate representations with cross-coordinate joint constraints. These designs enable Argus to achieve state-of-the-art overall performance among feed-forward methods on camera pose estimation, metric depth prediction, and point map reconstruction, while maintaining favorable efficiency. Argus excels in structured indoor environments but has clear limitations: its indoor-focused training data constrains zero-shot generalization to unbounded outdoor and aerial scenes, and memory scaling limits reconstruction from very large panorama sets (hundreds to thousands of views), making scale-agnostic reconstruction a promising future direction. More discussions are provided in the supplementary material. We believe Realsee3D and Argus together provide the community with a strong foundation for advancing metric panoramic 3D reconstruction and its downstream applications.

References

1. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M., Szeliski, R.: Building rome in a day. In: ICCV (2009)
2. Ai, H., Cao, Z., Wang, L.: A survey of representation learning, optimization strategies, and applications for omnidirectional vision: H. ai et al. IJCV **133**(8), 4973–5012 (2025)
3. Alama, O., Bhattacharya, A., He, H., Kim, S., Qiu, Y., Wang, W., Ho, C., Keetha, N., Scherer, S.: Rayfronts: Open-set semantic ray frontiers for online scene understanding and exploration. In: IROS. pp. 5930–5937 (2025)
4. Antequera, M.L., Gargallo, P., Hofinger, M., Bulo, S.R., Kuang, Y., Kontschieder, P.: Mapillary planet-scale depth dataset. In: ECCV. pp. 589–604 (2020)
5. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017)
6. Barath, D., Matas, J.: Graph-cut ransac. In: CVPR. pp. 6733–6741 (2018)
7. Besl, P.J., McKay, N.D.: Method for registration of 3-d shapes. In: Sensor fusion IV: control paradigms and data structures. vol. 1611, pp. 586–606. Spie (1992)
8. Bian, J., Lin, W.Y., Matsushita, Y., Yeung, S.K., Nguyen, T.D., Cheng, M.M.: Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In: CVPR. pp. 4181–4190 (2017)
9. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020)
10. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics (2021)
11. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: CVPR (2025)
12. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 3DV (2017)
13. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645 (2025)
14. Cruz, S., Hutchcroft, W., Li, Y., Khosravan, N., Boyadzhiev, I., Kang, S.B.: Zillow indoor dataset: Annotated floor plans with 360deg panoramas and 3d room layouts. In: CVPR. pp. 2133–2143 (2021)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
16. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ToG **36**(4), 1 (2017)
17. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. In: NeurIPS (2022)
18. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
19. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: CVPR. pp. 13142–13153 (2023)
20. Deng, J., Li, H., Xie, T., Ren, W., Zhang, Q., Tan, P., Guo, X.: Sail-recon: Large sfm by augmenting scene regression with localization. arXiv preprint arXiv:2508.17972 (2025)

21. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: CVPRW (2018)
22. Dijkstra, E.: A note on two problems in connexion with graphs. Numerische Mathematik **1**(1), 269–271 (1959)
23. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: 2025 International Conference on 3D Vision (3DV). pp. 1–10 (2025)
24. Edstedt, J., Athanasiadis, I., Wadenbäck, M., Felsberg, M.: Dkm: Dense kernelized feature matching for geometry estimation. In: CVPR (2023)
25. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: CVPR (2024)
26. Fei, X., Zheng, W., Duan, Y., Zhan, W., Tomizuka, M., Keutzer, K., Lu, J.: Driv3r: Learning dense 4d reconstruction for autonomous driving. arXiv preprint arXiv:2412.06777 (2024)
27. Feng, W., Qin, H., Wu, M., Yang, C., Li, Y., Li, X., An, Z., Huang, L., Zhang, Y., Magno, M., et al.: Quantized visual geometry grounded transformer. arXiv preprint arXiv:2509.21302 (2025)
28. Furukawa, Y., Hernández, C.: Multi-view stereo: A tutorial. Foundations and Trends in Computer Graphics and Vision **9**(1-2), 1–148 (2015)
29. Goesele, M., Curless, B., Seitz, S.M.: Multi-view stereo revisited. In: CVPR. vol. 2, pp. 2402–2409 (2006)
30. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: CVPR. pp. 3749–3761 (2022)
31. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth any camera: Zero-shot metric depth estimation from any camera. In: CVPR (2025)
32. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
33. Hausler, S., Garg, S., Xu, M., Milford, M., Fischer, T.: Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In: CVPR (2021)
34. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. In: CVPR (2024)
35. Hu, Y., Xie, Q., Jain, V., Francis, J., Patrikar, J., Keetha, N., Kim, S., Xie, Y., Zhang, T., Fang, H.S., et al.: Toward general-purpose robots via foundation models: A survey and meta-analysis. arXiv preprint arXiv:2312.08782 (2023)
36. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR. pp. 2821–2830 (2018)
37. Jang, W., Weinzaepfel, P., Leroy, V., Agapito, L., Revaud, J.: Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In: CVPR. pp. 1071–1081 (2025)
38. Jiang, H., Song, Z., Lou, Z., Xu, R., Tan, M.: Depth anything in 360^{circ}: Towards scale invariance in the wild. arXiv preprint arXiv:2512.22819 (2025)
39. Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al.: Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. TOG **44**(6), 1–16 (2025)
40. Jiang, Y., Yu, C., Xie, T., Li, X., Feng, Y., Wang, H., Li, M., Lau, H., Gao, F., Yang, Y., Jiang, C.: Vr-gs: A physical dynamics-aware interactive gaussian splatting system in virtual reality. In: ACM SIGGRAPH (2024)
41. Jung, D., Choi, J., Lee, Y., Manocha, D.: Im360: Large-scale indoor mapping with 360 cameras. In: ICCV (2025)

42. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: ICCV. pp. 6013–6022 (2025)
43. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: ECCV. pp. 18–35 (2024)
44. Keetha, N., Karhade, J., Jatavallabhula, K.M., Yang, G., Scherer, S., Ramanan, D., Luiten, J.: Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In: CVPR. pp. 21357–21366 (2024)
45. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
46. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. TOG (2023)
47. Kümmerle, C., Verdun, C.M., Stöger, D.: Iteratively reweighted least squares for basis pursuit with global linear convergence rate. In: NeurIPS. pp. 2873–2886 (2021)
48. Li, H., Zheng, W., He, J., Liu, Y., Lin, X., Yang, X., Chen, Y.C., Guo, C.: Da ^{2}: Depth anything in any direction. arXiv preprint arXiv:2509.26618 (2025)
49. Li, X.: Multi-view canonical pose 3d human body reconstruction based on volumetric tsdf. In: ECCV. pp. 392–397 (2022)
50. Li, X., Rao, T., Pan, C.: Edm: Efficient deep feature matching. In: ICCV. pp. 26198–26208 (2025)
51. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018)
52. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
53. Lin, X., Song, M., Zhang, D., Lu, W., Li, H., Du, B., Yang, M.H., Nguyen, T., Qi, L.: Depth any panoramas: A foundation model for panoramic depth estimation. arXiv preprint arXiv:2512.16913 (2025)
54. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: ICCV (2023)
55. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: CVPR. pp. 22160–22169 (2024)
56. Liu, Y., Min, Z., Wang, Z., Wu, J., Wang, T., Yuan, Y., Luo, Y., Guo, C.: Worldmirror: Universal 3d world reconstruction with any-prior prompting. arXiv preprint arXiv:2510.10726 (2025)
57. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
58. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
59. Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., Fan, X.: Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In: ICCV. pp. 6851–6860 (2019)
60. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

61. Moulon, P., Monasse, P., Perrot, R., Marlet, R.: Openmvg: Open multiple view geometry. In: International Workshop on Reproducible Research in Pattern Recognition. pp. 60–74. Springer (2016)
62. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE Transactions on Robotics (2015)
63. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
64. Pan, X., Charron, N., Yang, Y., Peters, S., Whelan, T., Kong, C., Parkhi, O., Newcombe, R., Ren, Y.C.: Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In: ICCV. pp. 20133–20143 (2023)
65. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: Unik3d: Universal camera monocular 3d estimation. In: CVPR (2025)
66. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
67. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: ICCV. pp. 10901–10911 (2021)
68. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV. pp. 10912–10922 (2021)
69. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: ICCV (2011)
70. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: CVPR (2020)
71. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016)
72. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: ECCV. pp. 501–518 (2016)
73. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
74. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A comparison and evaluation of multi-view stereo reconstruction algorithms. In: CVPR. vol. 1, pp. 519–528 (2006)
75. Shen, Y., Zhang, Z., Qu, Y., Cao, L.: Fastvggt: Training-free acceleration of visual geometry transformer. arXiv preprint arXiv:2509.02560 (2025)
76. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
77. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: CVPR (2021)
78. Sun, W., Jiang, W., Trulls, E., Tagliasacchi, A., Yi, K.M.: Acne: Attentive context normalization for robust permutation-equivariant learning. In: CVPR. pp. 11286–11295 (2020)
79. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D.S., Maksymets, O., et al.: Habitat 2.0: Training home assistants to rearrange their habitat. NeurIPS **34**, 251–266 (2021)
80. Taira, H., Okutomi, M., Sattler, T., Cimpoi, M., Pollefeys, M., Sivic, J., Pajdla, T., Torii, A.: Inloc: Indoor visual localization with dense matching and view synthesis. In: CVPR (2018)

81. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a modern synthesis. In: International workshop on vision algorithms (1999)
82. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. PAMI **13**(4), 376–380 (1991)
83. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
84. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
85. Wang, J., Chen, M., Zhang, S., Karaev, N., Schönberger, J., Labatut, P., Bojanowski, P., Novotny, D., Vedaldi, A., Rupprecht, C.: Vggt-ω. arXiv preprint arXiv:2605.15195 (2026)
86. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: CVPR (2024)
87. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR (2025)
88. 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)
89. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
90. Wang, Y., He, X., Peng, S., Tan, D., Zhou, X.: Efficient loftr: Semi-dense local feature matching with sparse-like speed. In: CVPR (2024)
91. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. In: ICLR (2026)
92. Wang, Y., Zeng, Z., Guan, T., Yang, W., Chen, Z., Liu, W., Xu, L., Luo, Y.: Adaptive patch deformation for textureless-resilient multi-view stereo. In: CVPR. pp. 1621–1630 (2023)
93. Xia, H., Fu, Y., Liu, S., Wang, X.: Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In: CVPR. pp. 22378–22389 (2024)
94. Yan, Z., Li, X., Wang, K., Zhang, Z., Li, J., Yang, J.: Multi-modal masked pretraining for monocular panoramic depth completion. In: ECCV. pp. 378–395 (2022)
95. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In: CVPR. pp. 21924–21935 (2025)
96. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
97. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS. pp. 21875–21911 (2024)
98. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: ECCV (2018)
99. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR. pp. 1790–1799 (2020)
100. Zhang, S., Wang, J., Xu, Y., Xue, N., Rupprecht, C., Zhou, X., Shen, Y., Wetzstein, G.: Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In: CVPR. pp. 21936–21947 (2025)

101. Zhao, Q., Lin, A., Tan, J., Zhang, J.Y., Ramanan, D., Tulsiani, S.: Diffusionsfm: Predicting structure and motion via ray origin and endpoint diffusion. In: CVPR (2025)
102. Zhao, X., Wu, X., Miao, J., Chen, W., Chen, P.C., Li, Z.: Alike: Accurate and lightweight keypoint detection and descriptor extraction. TMM pp. 3101–3112 (2022)
103. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: ECCV. pp. 519–535 (2020)
104. Zheng, X., Liao, C., Weng, Z., Lei, K., Dongfang, Z., He, H., Lyu, Y., Jiang, L., Qi, L., Chen, L., et al.: Panorama: The rise of omnidirectional vision in the embodied ai era. arXiv preprint arXiv:2509.12989 (2025)
105. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: ICCV. pp. 19855–19865 (2023)
106. Zhong, D., Zheng, X., Liao, C., Lyu, Y., Chen, J., Wu, S., Zhang, L., Hu, X.: Omnisam: Omnidirectional segment anything model for uda in panoramic semantic segmentation. In: ICCV. pp. 23892–23901 (2025)
107. Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., Lu, J.: Streaming 4d visual geometry transformer. arXiv preprint arXiv:2507.11539 (2025)