MapAnything: Universal Feed-Forward Metric 3D Reconstruction

map-anything.github.io

Nikhil Keetha [1,2] Norman Müller [1] Johannes Schönberger [1] Lorenzo Porzi [1] Yuchen Zhang [2]
Tobias Fischer [1] Arno Knapitsch [1] Duncan Zauss [1] Ethan Weber [1] Nelson Antunes [1]
Jonathon Luiten [1] Manuel Lopez-Antequera [1] Samuel Rota Bulò [1] Christian Richardt [1]
Deva Ramanan [2] Sebastian Scherer [2] Peter Kontschieder [1]

[1] Meta Reality Labs [2] Carnegie Mellon University

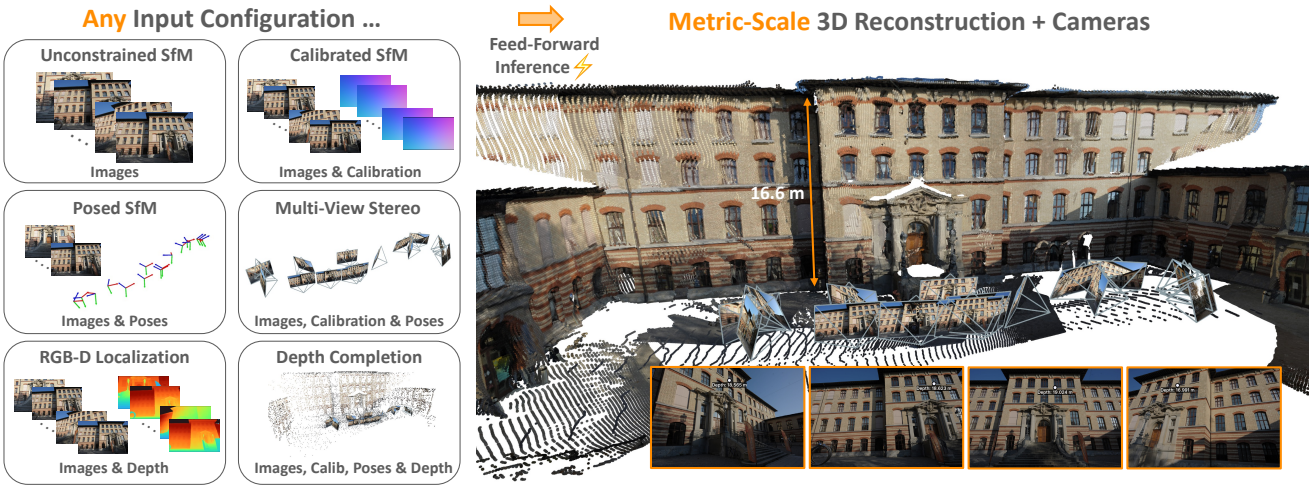

Figure 1. **MapAnything is a flexible, unified feed-forward 3D reconstruction model** that predicts metric 3D reconstructions with camera information from a set of N input images with optional camera poses, intrinsics, or depth maps. MapAnything supports over 12 different 3D reconstruction tasks, including camera localization, structure-from-motion (SfM), multi-view stereo, and metric depth completion, outperforming or matching the quality of specialist methods.

Abstract

We introduce MapAnything, a unified transformer-based feed-forward model that ingests one or more images along with optional geometric inputs such as camera intrinsics, poses, depth, or partial reconstructions, and then directly regresses the metric 3D scene geometry and cameras. MapAnything leverages a factored representation of multi-view scene geometry, i.e., a collection of depth maps, local ray maps, camera poses, and a metric scale factor that effectively upgrades local reconstructions into a globally consistent metric frame. Standardizing the supervision and training across diverse datasets, along with flexible input augmentation, enables MapAnything to address a broad range of 3D vision tasks in a single feed-forward pass, including uncalibrated structure-from-motion, calibrated multi-view stereo, monocular depth estimation, camera localization, depth completion, and more. We provide extensive experimental analyses and model ablations demonstrating that MapAnything outperforms or matches specialist feed-forward models while offering more efficient joint training behavior, thus paving the way toward a universal 3D reconstruction backbone.

1. Introduction

The problem of image-based 3D reconstruction has traditionally been solved using structure-from-motion (SfM) [15, 18], photometric stereo [28], shape-from-shading [7], and so on. To make the problem tractable, classic approaches decompose it into distinct tasks, such as feature detection [13] and matching [16], two-view pose estimation [14], camera calibration [23] and resectioning [17], rotation [4] and translation averaging [15], bundle adjustment (BA) [21], multi-view stereo (MVS) [19], and/or monocular surface estimation [6]. Recent work has demonstrated tremendous potential in solving these problems in a unified way using feed-forward architectures [2, 9, 12, 25, 26, 30].

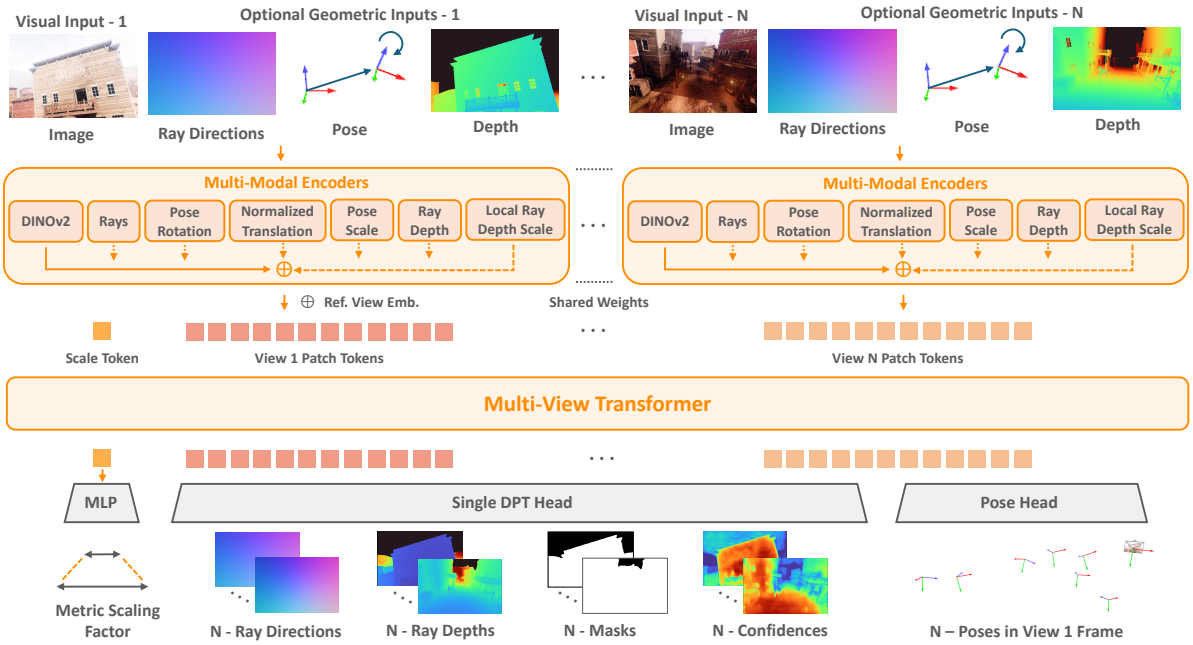

Figure 2. **Overview of the MapAnything Architecture.** Given N visual and optional geometric inputs, the model first encodes the images and the factored representation of the geometric inputs into a common latent space where the patch features (for images, rays & depth) and broadcasted global features (for translation, rotation, pose scale across all pose inputs & depth scale local to each frame) are summed together. Then, a fixed reference view embedding is added to the first view's features and a single learnable scale token is appended to the set of N view patch tokens. These tokens are then input into an alternating-attention transformer. We use a single DPT to decode the N view patch tokens into N dense outputs local to all the views. A single average pooling-based pose head also uses the N view patch tokens to predict N poses in the frame of view 1. Lastly, while these predictions exist in an up-to-scale space, the model passes the scale token through an MLP to predict the metric scaling factor, which when coupled with the other predictions, provides the dense metric 3D reconstruction.

While prior feed-forward work has approached the different tasks disjointly or by not leveraging all the available input modalities, we present a unified end-to-end model for diverse 3D reconstruction tasks. Our method MapAnything can be used to solve the most general uncalibrated SfM problem as well as various combinations of sub-problems, like calibrated SfM or multi-view stereo, monocular depth estimation, camera localization, metric depth completion, etc. To enable the training of such a unified model, we: (1) introduce a flexible input scheme that supports various geometric modalities when available, (2) propose a suitable output space that supports all of these diverse tasks, and (3) discuss flexible dataset aggregation and standardization.

MapAnything's key insight to address these challenges is the use of a *factored* representation of multi-view scene geometry. Instead of directly representing the scene as a collection of pointmaps, we represent the scene as a collection of depthmaps, local raymaps, camera poses, and a metric scale factor that upgrade local reconstructions into a globally consistent metric frame. We use such a factored representation to represent both the outputs and (optional) inputs for MapAnything, allowing it to take advantage of auxiliary geometric inputs when available. For example, robotic applications [1, 5, 8, 11] may have knowledge of camera intrinsics (rays) and/or extrinsics (pose). Finally, a significant benefit of our factored representation is that it allows for MapAnything to be effectively trained from diverse datasets with partial annotations, for example, datasets that may be annotated with only non-metric "up-to-scale" geometry. In summary, we make the following main contributions:

1. **Unified Feed-Forward Model** for multi-view metric 3D reconstruction that supports more than 12 different problem configurations. The end-to-end transformer is trained more efficiently than a naive set of bespoke models and leverages not only image inputs, but also optional geometric information such as camera intrinsics, extrinsics, depth, and/or metric scale factor, when available.
2. **Factored Scene Representation** that flexibly enables decoupled inputs and effective prediction of metric 3D reconstructions. Our model computes multi-view pixel-wise scene geometry and cameras directly, without redundancies or costly post-processing.
3. **State-of-the-Art Performance** compared to other feed-forward models, matching or surpassing expert models that are tailored for specific, isolated tasks.
4. **Open Source Release** of (a) code for data processing, inference, benchmarking, training & ablations, and (b)

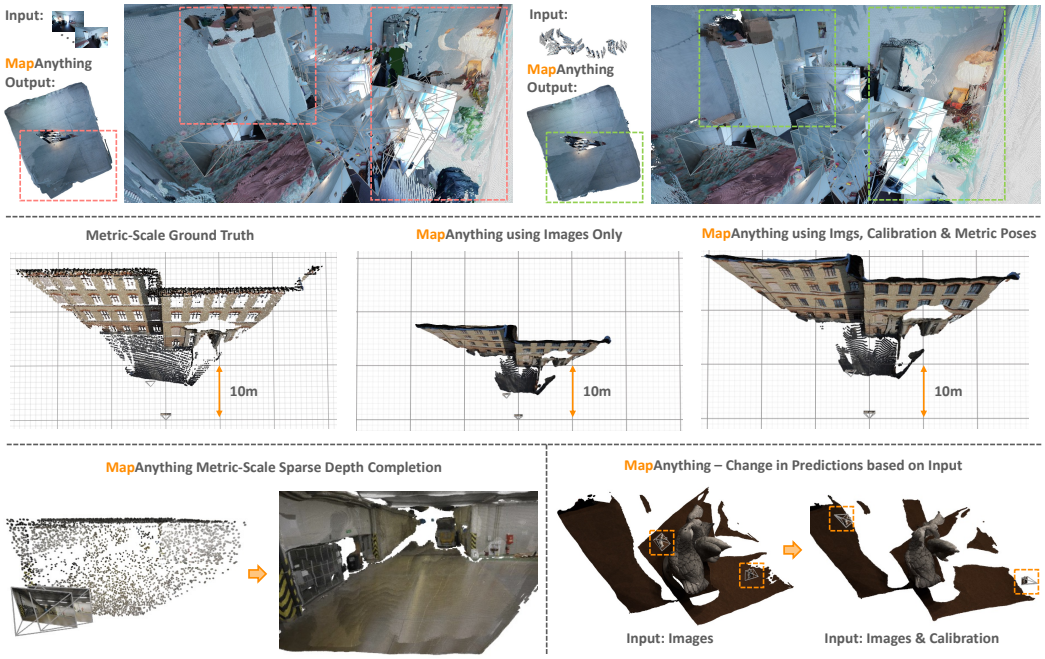

Figure 3. **Auxiliary geometric inputs improve feed-forward performance of MapAnything.** `(Top)` While MapAnything & other baselines using 100 input images show duplication of 3D structure, when provided with the camera calibration and poses, the 3D reconstruction significantly improves, showcasing aligned geometry. `(Middle)` MapAnything using images only as input showcases non-precise metric scale estimation on ETH3D (a zero-shot dataset). However, when the calibration and metric poses are provided as additional input, the estimated metric scale significantly improves and approximately matches the ground truth. `(Bottom-Left)` We showcase that MapAnything is able to leverage a sparse metric pointcloud as input to perform dense metric depth completion. `(Bottom-Right)` Despite not being trained for object-centric data, we showcase how the scene geometry and cameras change based on the input provided.

a pre-trained MapAnything model under the permissive Apache 2.0 license, thereby providing an extensible & modular framework plus model to facilitate future research on building 3D/4D foundation models.

2. MapAnything

MapAnything is an end-to-end model that takes as input N RGB images $\hat{\mathcal{I}} = (\hat{I}_i)_{i=1}^N$ and optional geometric inputs corresponding to all or a subset of the input views:

- **(a)** generic central camera calibrations [3, 22, 30] as ray directions $\hat{\mathcal{R}} = (\hat{R}_i)_{i \in S_r}$,
- **(b)** poses in the frame of the first view $\hat{I}_1$ as quaternions $\hat{\mathcal{Q}} = (\hat{Q}_i)_{i \in S_q}$ and translations $\hat{\mathcal{T}} = (\hat{T}_i)_{i \in S_t}$, and
- **(c)** ray depth for each pixel $\hat{\mathcal{D}} = (\hat{D}_i)_{i \in S_d}$,

where S_r, S_q, S_t, S_d are subsets of frame indices $[1, N]$.

MapAnything maps these inputs to an N-view factored metric 3D output (as shown in Figure 2):

$$f_{\text{MapAnything}}\big(\hat{\mathcal{I}}, [\hat{\mathcal{R}}, \hat{\mathcal{Q}}, \hat{\mathcal{T}}, \hat{\mathcal{D}}]\big) = \{m, (R_i, \tilde{D}_i, \tilde{P}_i)_{i=1}^N)\}, \quad (1)$$

where $m \in \mathbb{R}$ is the predicted global metric scaling factor, and for each view i, $R_i \in \mathbb{R}^{3\times H\times W}$ are the predicted local ray directions, $\tilde{D}_i \in \mathbb{R}^{1\times H\times W}$ are the ray depths in a up-to-scale space (indicated by the tilde), and $\tilde{P}_i \in \mathbb{R}^{4\times 4}$ is the pose of image $\hat{I}_i$ in the frame of image $\hat{I}_1$, represented as quaternion $Q_i \in \mathrm{SU}(2)$ and up-to-scale translation $\tilde{T}_i \in \mathbb{R}^3$.

We can further use this factored output to get the up-to-scale local point maps (3D points corresponding to each pixel) as $\tilde{L}_i = R_i \cdot \tilde{D}_i \in \mathbb{R}^{3\times H\times W}$. Then, leveraging the rotation matrix $O_i \in \mathrm{SO}(3)$ (obtained from Q_i) and up-to-scale translation, we can compute the N-view up-to-scale point maps in world frame as $\tilde{X}_i = O_i \cdot \tilde{L}_i + T_i$. The final metric 3D reconstruction for the N input views (in the frame of image 1) is given by $X_i^{\text{metric}} = m \cdot \tilde{X}_i$ for $i \in [1, N]$. Please see our full paper on the MapAnything webiste for more details.

3. Benchmarking & Results

In this section, we benchmark MapAnything against expert baselines specifically designed or trained for the task. We perform all the experiments with a constant seed. Please see our paper PDF for the full suite of 3D vision benchmarking results, ranging from unconstrained SfM, multi-view stereo (MVS), calibration to localization, and depth completion. In the full PDF, we also provide ablations providing key insights into enabling MapAnything.

Multi-View Dense Reconstruction: We benchmark the performance of pointmaps, pose, depth & ray directions estimation on an undistorted version of ETH3D [20], ScanNet++ v2 [29], and TartanAirV2-WB [27, 31], where, for each test scene, we randomly sample up to N views that form a sin-

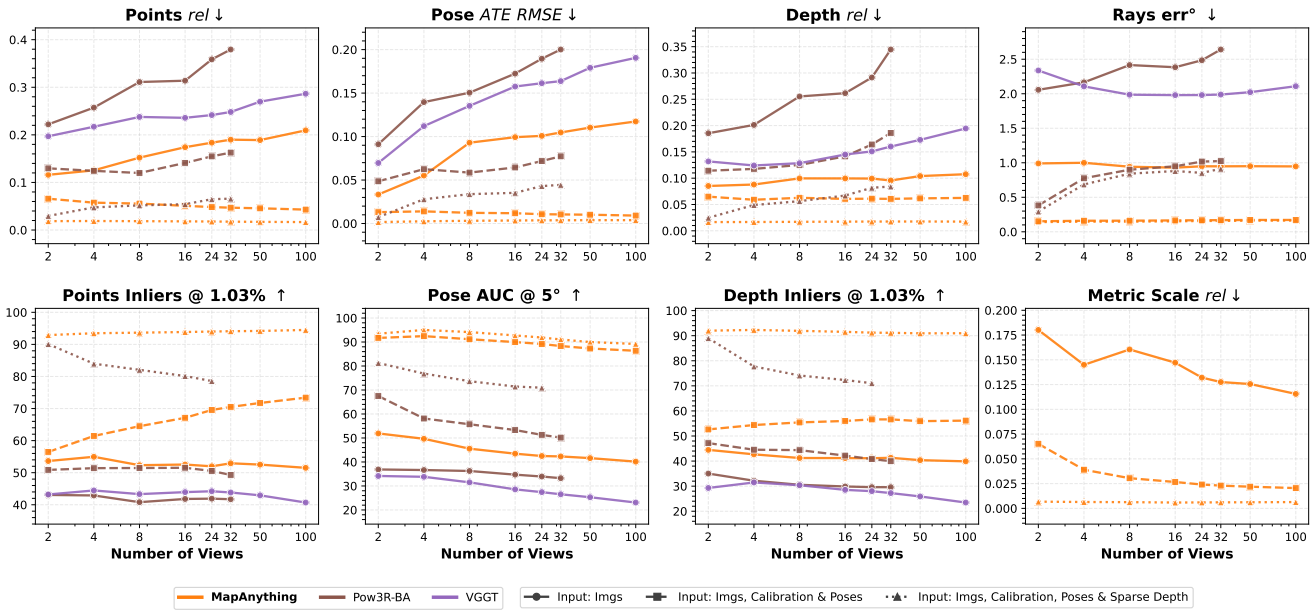

Figure 4. **MapAnything shows state-of-the-art dense multi-view reconstruction for number of input views varying from 2 to 100 and under different input configurations.** We report the absolute relative error (rel), the inlier ratio at a relative threshold of 1.03% (τ), the average aligned trajectory error ($ATE\ RMSE$), the area under the curve at an error threshold of 5° ($AUC@5$), and the average angular error (err) in degrees (°), averaged over ETH3D, ScanNet++ v2 & TAv2. We don't report performance for baselines when the inference runs out of GPU memory. We provide results for the exhaustive input configurations of MapAnything in the supplement.

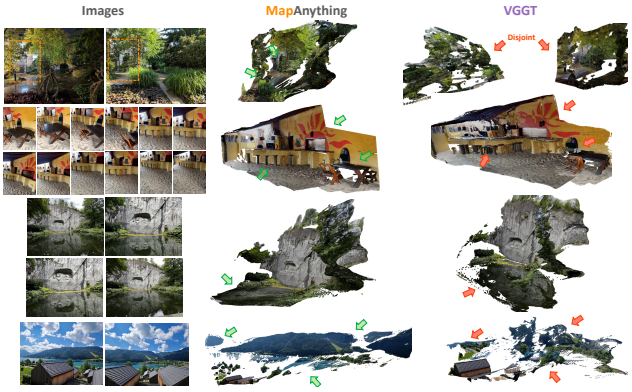

Figure 5. **Qualitative comparison of MapAnything to VGGT [24] using only in-the-wild images as input.** For a fair comparison, we apply the same normal-based edge mask post-processing and our sky mask to both methods. MapAnything more effectively deals with large disparity changes, seasonal shifts, textureless surfaces, water bodies and large scenes.

gle connected component graph based on the pre-computed pair-wise covisibility of all images in the scene (this prevents disjoint sets of images as input). Figure 4 shows that MapAnything provides state-of-the-art dense multi-view reconstruction performance over other baselines using only image input, including VGGT [24]. Beyond the performance using only images as input, we show that MapAnything can leverage additional auxiliary geometric inputs for feed-forward inference to further increase reconstruction performance by a significant factor. Furthermore, we find MapAnything is better than the bundle adjustment (BA) variant of the two-view baseline, Pow3R [10], which is also designed to leverage scene priors. In Figure 3, we illustrate how the auxiliary geometric inputs improve MapAnything. We also find that the reconstruction outputs from MapAnything (using only images as input) display high fidelity, as shown in Figure 5.

4. Conclusion

MapAnything is the first universal transformer-based backbone that directly regresses metric 3D geometry and camera poses from flexible inputs – including images, camera intrinsics, poses, depth maps, or partial reconstructions – in a single pass. By using a factored representation of multi-view geometry (depth maps, ray maps, poses, and a global scale factor), MapAnything unifies local estimates into a global metric frame. With standardized supervision across varied datasets and augmentations, MapAnything handles multiple tasks like uncalibrated structure-from-motion, calibrated multi-view stereo, monocular depth estimation, camera localization, depth completion, and more without task-specific tuning. Extensive experiments show that it surpasses or matches specialist models while enabling efficient joint training. Future extensions to dynamic scenes, uncertainty quantification, and scene understanding promise to further generalize MapAnything's capabilities and robustness, paving the way toward a truly universal 3D reconstruction backbone.

References

[1] Omar Alama, Avigyan Bhattacharya, Haoyang He, Seungchan Kim, Yuheng Qiu, Wenshan Wang, Cherie Ho, Nikhil Keetha, and Sebastian Scherer. Rayfronts: Open-set semantic ray frontiers for online scene understanding and exploration. In *IROS*, 2025. 2

[2] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *CVPR*, 2025. 1

[3] Michael D Grossberg and Shree K Nayar. A general imaging model and a method for finding its parameters. In *ICCV*, 2001. 3

[4] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *IJCV*, 103:267–305, 2013. 1

[5] Cherie Ho, Jiaye Zou, Omar Alama, Sai M Kumar, Benjamin Chiang, Taneesh Gupta, Chen Wang, Nikhil Keetha, Katia Sycara, and Sebastian Scherer. Map it anywhere: Empowering bev map prediction using large-scale public datasets. In *NeurIPS*, 2024. 2

[6] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric context from a single image. In *ICCV*, 2005. 1

[7] Berthold KP Horn. Obtaining shape from shading information. In *Shape from shading*, pages 123–171. MIT Press, 1989. 1

[8] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Hao-Shu Fang, et al. Toward general-purpose robots via foundation models: A survey and meta-analysis. *arXiv preprint arXiv:2312.08782*, 2023. 2

[9] Sergio Izquierdo, Mohamed Sayed, Michael Firman, Guillermo Garcia-Hernando, Daniyar Turmukhambetov, Javier Civera, Oisin Mac Aodha, Gabriel Brostow, and Jamie Watson. MVSAnywhere: Zero-shot multi-view stereo. In *CVPR*, 2025. 1

[10] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3R: Empowering unconstrained 3D reconstruction with camera and scene priors. In *CVPR*, 2025. 4

[11] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *CVPR*, 2024. 2

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *ECCV*, 2024. 1

[13] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 1

[14] David Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004. 1

[15] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *ECCV*, 2024. 1

[16] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1

[17] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2D-to-3D matching. In *ICCV*, 2011. 1

[18] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[19] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 1

[20] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 3

[21] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International Workshop on Vision Algorithms*, 2000. 1

[22] Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural ray surfaces for self-supervised learning of depth and ego-motion. In *3DV*, 2020. 3

[23] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image calibration with geometric optimization. In *ECCV*, 2024. 1

[24] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 4

[25] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. PF-LRM: Pose-free large reconstruction model for joint pose and shape prediction. In *ICLR*, 2024. 1

[26] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, 2024. 1

[27] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual SLAM. In *IROS*, 2020. 3

[28] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical Engineering*, 19(1):139–144, 1980. 1

[29] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, 2023. 3

[30] Jason Y. Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *ICLR*, 2024. 1, 3

[31] Yuchen Zhang, Nikhil Keetha, Chenwei Lyu, Bhuvan Jhamb, Yutian Chen, Yuheng Qiu, Jay Karhade, Shreyas Jha, Yaoyu Hu, Deva Ramanan, Sebastian Scherer, and Wenshan Wang. UFM: A simple path towards unified dense correspondence with flow. arXiv:2506.09278, 2025. 3