AMB3R: Accurate Feed-forward Metric-scale 3D Reconstruction with Backend

Supplementary Material

7. Limitations

7.1. Scene Representations

Local scene representation. Our backend representation is a local scene representation that cannot perform joint reasoning across multiple chunks predicted by the front-end. Extending it to a global representation can be a promising direction, as it would enable long-term consistency across chunks. However, training such a global backend would also require substantially more computational resources.

Computational complexity. Our model shares the inherent limitation of type-(c) methods regarding the quadratic computational complexity with respect to the number of input images. However, a key advantage of our sparse yet compact 3D representation is that its complexity depends on the amount of 3D content rather than the number of views. This is, in fact, one of our motivations to introduce a compact 3D scene representation as a backend. Due to the computational constraint, we are not able to investigate on how to re-design the entire pointmap-based foundation models. However, our work serves as a proof-of-concept demonstrating the effectiveness of the spatial compactness. We believe this opens a promising direction towards making pointmap-based foundation models themselves more scalable to longer sequences. One possible direction is to reduce the usage of global attention and leverage the compact 3D scene representation as an alternative to reduce the computational complexity for long sequences.

Dynamic scenes. Since our model has not been trained for dynamic scenarios, it heavily relies on static cues for dynamic environment. Thus, it might fail when the target scene is dominated by dynamic objects. This can be solved by including dynamic scenes as training datasets.

7.2. Visual Odometry

Reliance on dense reconstruction prior. Our visual odometry relies on the prior that the model predicts geometry in reference (first) frame's coordinate system, up to an unknown scale factor. When this assumption breaks, typically in scenes with diverse depth ranges and complex thin structures, the scale alignment becomes unreliable, often leading to tracking failure. Similarly, scenes dominated by distant content can increase prediction variance, resulting in poor scale alignment and trajectory accuracy. In this case, type-(b) method might benefit from its persistent memory and a generally longer perception window.

Drift/Kidnapping issue. Our system is a visual odometry without explicit loop closure or relocalization. Consequently, tracking might drift in large-scale environments or fail under long-term kidnapping scenarios. In this case, an optimization-based module could help for loop closure or relocalization. Notably, our backward search strategy in keyframe selection could also mitigate this issue. Once the loop is closed, as long as the accumulative pose error is less than the backward search pose distance, our model can still leverage the earlier frame to bring the system back on track. This is useful for an online system, as it is not necessary to fix the earlier trajectory in that case.

7.3. Structure from Motion

Reliance on dense reconstruction prior. Our SfM shares the same limitations as in VO as they rely on the same prior.

Initialization. Our initialization is purely based on the feature-based image clustering and the confidence of the local chunk. We notice that our confidence usually prefers indoor regions. In some photo-tourism scenarios, our model might initialize with indoor images that do not guarantee overlap with the outdoor regions that are of primary interest. One possible solution is to use explicit feature matching to construct the view graph at cost of extra complexity.

Mapping window selection. Our mapping window selection for global mapping is based on pose distance. This is usually sufficient for visual odometry due to the local smoothness of the trajectory. However, for a large-scale structure from motion system, the same geometry is likely to be observed from diverse views. Selecting mapping frames using pose distance can therefore omit overlapping views with wide baselines. In that case, a geometry-based view selection might help at the cost of extra complexity.

Non-overlap reconstruction. The ability to reconstruct non-overlapping views is useful for small-scale problems where overlap genuinely does not exist. However, this could become problematic at larger scales: image clustering may group visually similar images from entirely different locations into the same cluster despite having no geometric overlap. In such cases, the model may hallucinate geometry for these non-overlapping images and give them moderate confidence, preventing them from being pruned and leading to irreversible reconstruction errors. One possible remedy is to filter out false clusters (e.g., Doppelgangers [14]).

Pose consistency. The final camera poses are weighted average feed-forward poses. Compared to COLMAP [101] poses obtained via BA, our poses may not strictly follow the geometry constraint. For downstream tasks like novel view synthesis, which requires strict pose coherence, our poses, even metrically better in some cases, might still result in lower PSNR. This is the inherent limitation of feed-forward poses and can be addressed via BA as post-processing.

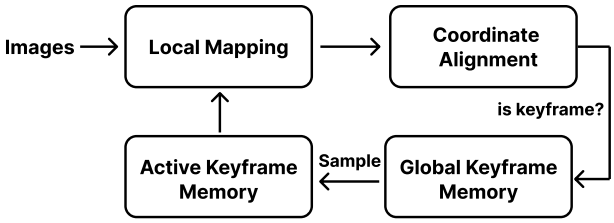

Figure 4. **Overview of AMB3R (VO).** Input frames are mapped with the keyframes in the active keyframe memory to predict geometry and camera poses. After coordinate alignment, we select new keyframes and update the global keyframe memory; poses and geometry for non-keyframes are also stored. If the active keyframe memory is not full, the new keyframe is appended; otherwise, we refresh the active keyframe memory by resampling a new set of keyframes from the global keyframe memory.

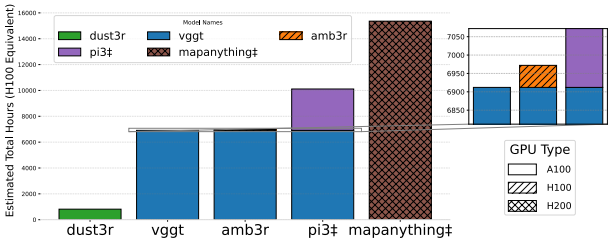

Figure 5. **Training cost comparison.** We roughly estimate the training cost of each model. We count 1 H100/H200 hours (AMB3R/MapAnything) as 2 A100 hours.

8. Additional Details

8.1. Metric-scale Head

To predict the metric scale factor, we first map the encoder feature into one feature vector via a three-layer MLP. We then add this feature with the depth feature from the depth DPT [96] branch of VGGT [126], and map this new feature into a metric scale factor via a two-layer convolution. We supervise this prediction using an $\mathcal{L}_1$ loss. In cases where the selected pixels contain missing depth values, which frequently occurs in the Waymo [112] dataset, we use a ROE solver to estimate the relative scale difference. This allows us to recover a consistent ground-truth log-depth value for supervision even when raw depth is unavailable.

8.2. Training Datasets

We use ScanNet [22], ScannetPP [141], WildRGBD [136], Mapfree [4], Aria [91], Waymo [112], Virtual Kitti2 [12], GTASfM [127], MVS-Synth [44], OmniObject3d [132], and Hypersim [100] to train our model. Due to the constraint of the data storage, we only store a subset of each dataset. Although we only sample 2k samples in total for each epoch, we find data diversity is significantly more important than the data amount for training pointmap-based foundation model. We exclude Co3D [97] as we observe that VGGT might overfit on certain patterns on Co3D.

8.3. Training cost

We show the estimated training cost comparison in Fig. 5, where 1 H100/H200 hour counts as 2 A100 hours. Since π3 did not report their overall training time, we roughly estimate their training cost based the number of training epochs and GPUs used for training. The training of our backend and metric-scale head requires around 80 H100 hours in total. Compared to π3 and MapAnything, our model requires significantly less add-on cost on top of VGGT [126].

We note, however, that as pointed out by the MapAnything team, the training cost of MapAnything can be reduced to roughly match that of VGGT by reducing the total training data by half at the cost of a slight performance drop.[1]

8.4. Visual Odometry Pipeline

Fig. 4 illustrates our visual odometry pipeline. Input frames are mapped with keyframes stored in the active keyframe memory to predict camera poses and geometry. The coordinate alignment is done via 1) transforming the active keyframe map from global to local coordinates, 2) estimating the relative scale of the corresponding keyframe geometry, and 3) transforming the local map to global coordinates via the weighted average of relative poses of each corresponding keyframe. We then select new keyframes from the newly mapped frames and update the global keyframe memory. If the number of keyframes in active keyframe memory has not reached its capacity, we append the new keyframe; otherwise, we refresh the entire active keyframe memory by resampling from the global keyframe memory.

8.5. Structure from Motion Pipeline

We show the pipeline of our SfM as in Fig. 6. Our SfM mainly consists of 3 stages: 1) Image clustering that groups images into small clusters, 2) Coarse registration that registers each cluster incrementally, and 3) Global mapping that refines keyframes and non-keyframes via mapping.

8.6. Visual Odometry Runtime Analysis

We evaluate the runtime of our visual odometry on the TUM dataset [110]. Note that we exclude the data-loading overhead. Our method runs at an average of 4.2 FPS, with a best case of 6.0 FPS and a worst case of 3.4 FPS on an NVIDIA RTX 4090 GPU with input resolution of (392, 518). The variation in speed is primarily caused by the fluctuating number of active keyframes and the scene difficulty (The backend can be skipped if the front-end confidence

[1] https://github.com/HengyiWang/amb3r/issues/5

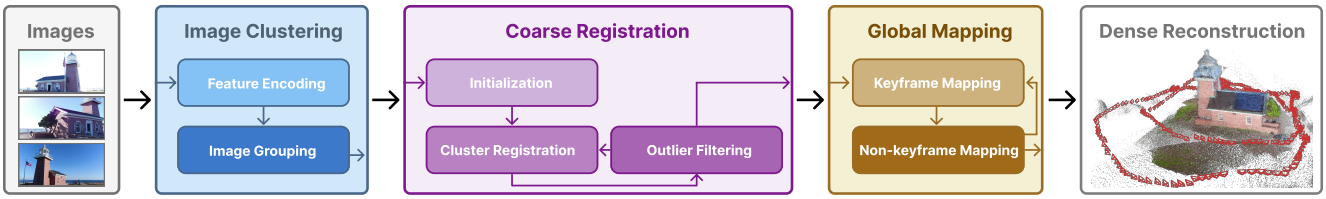

Figure 6. **Overview of AMB3R (SfM).** Our SfM pipeline contains 1) Image clustering that groups images into small clusters, 2) Coarse registration that constructs an initial coarse reconstruction, and 3) Global mapping that performs keyframe and non-keyframe refinement.

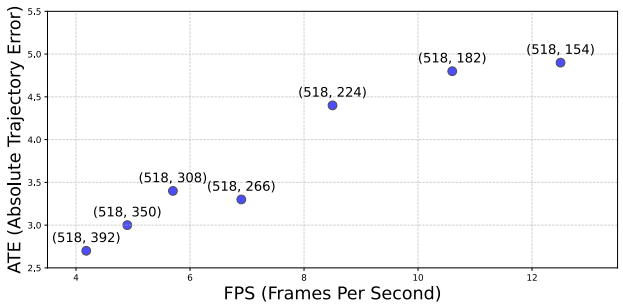

Figure 7. **VO runtime with respect to resolution.** By reducing input resolution, our method can achieve over 10FPS inference while being more accurate compared to other uncalibrated VO.

is sufficiently high). Importantly, since we cap the maximum number of active keyframes at 10, **the computational complexity does not grow with respect to the number of frames**, unlike type-(b) methods that do not have memory pruning.

Depending on the computation budget, one might reduce the input resolution or number of active keyframes for better runtime at the cost of accuracy (See Fig. 7).

8.7. Imperfect Ground-truth on 7 Scenes

We find that the pseudo ground truth in the original 7-Scenes [107] dataset often contains noticeable drift in certain sequences. This is because those poses are obtained via ICP-based KinectFusion [87]. ICP is prone to have rotation drift in scenes with near-spherical geometry and translation drift in scenes with many flat surfaces. The 7-scene dataset is the latter case. For instance, in Pumpkin/seq01, we find a substantial drift toward the end of the sequence. As a result, it is fundamentally impossible to achieve accurate tracking when evaluated against this pseudo GT unless the model happens to drift in exactly the same way. This makes the original pseudo GT unsuitable for fair comparison.

To address this, we adopt COLMAP GT [8] that is obtained via global optimization across all sequences of each scene. To ensure its reliability, we follow ACE-Zero [10] and evaluate novel view synthesis using NeRFstudio [114]. As shown in Tab. 7, the PSNR obtained from rendering with the new GT is consistently higher, confirming its accuracy. We therefore use it as the new ground-truth for evaluating 3D reconstruction and visual odometry.

Figure 8. **Examples** of buildings in DTU [2] dataset.

8.8. Baselines and Evaluation Datasets

Monocular depth estimation. We compare against Omnidata [29], Depth Anything v2 [139], Marigold [53], Diffusion-E2E [36], and MoGe [129], which are explicitly trained for monocular depth estimation. In contrast, VGGT [126] and our model are trained with a multi-view objective, and evaluated on monocular depth in a zero-shot manner. We use NYUv2 [86], KITTI [37], ETH3D [102], ScanNet [22], and DIODE [121] for evaluation following Marigold [53]. Note that for DIODE, there is a known issue about floaters in ground-truth geometry. Existing works like depth anything v2 [139] and MoGe [129] evaluate on DIODE with a pre-processing script that excludes those floaters. Since these scripts are not publicly available, we evaluate on the original noisy ground-truth following Marigold and Diffusion-E2E.

Multi-view depth estimation. We compare with 5 different categories of methods here: a) classic approach: COLMAP [101], which is an integrated solution of SfM and MVS based on optimization, b) monocular depth estimation methods that estimate depth from a single view: Depth Pro [6], Metric3D [142], UniDepthV2 [94], and Depth Anything v2 [139], c) Multi-view depth estimation methods that requires known calibration, camera poses, and per-image range: MVSNet [140], Vis-MVSNet [147], PatchmatchNet [123], and MVSFormer++ [15]. d) Multi-view depth estimation methods without the need for per-image range: Fast-MVSNet [143], Robust MVD baseline [104], and MVSA [46]. and e) Depth estimation from raw images without any prior information required: DeMoN [120], DUSt3R [130], Spann3R [124], Pow3R [48], MUSt3R [13], VGGT [126], as well as concurrent works π3 [131], and MapAnything [54]. We evaluate those methods on RMVDB [104], using KITTI [37], ETH3D [102], ScanNet [22], DTU [2], and Tanks & Temples [57] datasets.

Metric-scale estimation. We compare with existing

pointmap-based foundation models, which can recover metric-scale factors: MASt3R [60], Spann3R [124], MUSt3R [13], CUT3R [128], and concurrent work MapAnything [54] on RMVDB [104]. Note that DTU datasets contain many buildings placed on a pure white table as in Fig. 8. Due to the lack of background environment for scale reasoning, existing models would usually predict those buildings as real buildings, resulting in high absolute relative errors. In that case, the inlier ratio could be a good measure of the metric-scale prediction.

3D reconstruction. We compare with existing representative pointmap-based foundation models: Spann3R [124], MUSt3R [13], CUT3R [128], VGGT [126], as well as concurrent works π3 [131], and MapAnything [54] on ETH3D [102], DTU [2], and 7 scenes [107]. We use image tuples from RMVDB [104] for ETH3D and DTU. For 7 scenes, we use Spann3R [124] split with improved GT. Note that since pointmap prediction has an exact one-to-one correspondence with groundtruth, we remove ICP alignment in the prior evaluation protocol [124, 125].

Video depth estimation. We compare with pointmap-based models: Spann3R [124], CUT3R [128], VGGT [126], and π3 [131] on dynamic video depth estimation task on Sintel [11], Bonn [89], and Kitti [37] datasets. Among those methods, CUT3R [128] and π3 [131] are trained on dynamic datasets, while our model is not specifically trained on dynamic data.

VO and SLAM. For visual odometry, we compare with 1) Sparse VO: ORB-SLAM3 [84], DSO [31], and DPVO [69], 2) Dense VO: TANDEM [58], MonoGS [79], DeepFactors [21], DepthCov [25], DROID-VO [118], COMO [26], and GLORIE-VO [145], and 3) Uncalibrated VO: Spann3R [124] and MUSt3R [13]. In addition to visual odometry baselines, we also consider SLAM baselines with global bundle adjustment and loop closure: 1) SLAM with calibration: ORB-SLAM3 [84], DeepV2D [117], DeepFactors [21], DPV-SLAM [69], GO-SLAM [152], DROID-SLAM [118], MASt3R-SLAM [85] and 2) SLAM without calibration: DROID-SLAM [118], MASt3R-SLAM [85], VGGT-SLAM [126]. We compare with those methods using common SLAM benchmarks, including TUM [110], ETH-SLAM [103], and 7scenes [107]. We also compare with MegaSaM [64], a structure-and-motion method specifically designed for dynamic environments, on the TUM Dynamic [110] dataset to test the generalization on dynamic scenes.

Structure from Motion. We compare with optimization-based SfM methods: COLMAP [101], ACE-Zero [10], FlowMAP [108], VGGSfM [126], DF-SfM [39], and MASt3R-SfM [28] on ETH3D [102] and Tanks&Temples [57] dataset. ETH3D [102] contains unordered image collection, while Tanks&Temples [57] contains images from video.

Method	KITTI		ScanNet		ETH3D		DTU		T&T		Avg	
	rel ↓	δ ↑	rel ↓	δ ↑	rel ↓	δ ↑	rel ↓	δ ↑	rel ↓	δ ↑	rel ↓	δ ↑
w/o backend	4.5	60.4	(2.3)	(80.8)	1.8	85.3	1.0	94.8	2.0	83.9	2.3	80.6
w/ 2d backend	2.9	73.9	(2.0)	(84.2)	1.4	90.2	1.1	94.4	1.8	88.4	1.8	86.2
w/o scale align	3.0	72.0	(2.0)	(84.8)	1.5	89.7	**0.9**	**95.3**	1.9	89.0	1.9	86.2
w/o zero conv	26.9	7.8	(17.5)	(15.5)	19.8	17.8	7.3	34.6	16.2	23.9	17.5	19.9
Full	**2.8**	**74.4**	**(1.9)**	**(85.8)**	**1.4**	**90.9**	0.9	95.1	**1.7**	**90.2**	**1.7**	**87.3**

Table 17. **Ablation studies**. We ablate design choices of backend, scale alignment for supervision, and zero convolution.

Method	KITTI		ScanNet		ETH3D		DTU		T&T	
	rel	$\delta_{1.25}$	rel	$\delta_{1.25}$	rel	$\delta_{1.25}$	rel	$\delta_{1.25}$	rel	$\delta_{1.25}$
w/o backend	8.9	91.1	**(8.9)**	(95.2)	9.1	89.4	243.5	55.3	8.2	94.3
w/o decoder feature	10.2	82.9	(12.0)	(81.7)	**6.6**	**94.8**	**216.2**	52.7	7.6	94.1
Ours	**8.2**	**95.6**	(9.3)	**(95.2)**	8.5	90.4	240.9	**60.9**	**6.3**	**95.7**

Table 18. **Ablation study on metric-scale depth estimation**.

	Scene	# Imgs	RRA			RTA			mAA		
			@5	@15	@30	@5	@15	@30	@5	@10	@30
Training	Brandenburg Gate	1363	99.8	100.0	100.0	79.1	94.6	98.0	56.1	71.8	87.9
	Buckingham Palace	1676	88.0	88.2	88.3	86.6	92.8	95.0	69.8	77.0	83.6
	Colosseum Exterior	2063	77.7	79.3	79.5	73.8	85.4	88.7	46.1	58.9	70.9
	Grand Place Brussels	1083	58.0	59.6	60.9	62.8	73.3	78.2	37.1	44.0	51.7
	Notre Dame Front Facade	3765	20.2	28.0	40.4	37.1	45.4	60.7	13.6	15.5	19.3
	Palace of Westminster	983	78.4	78.7	79.1	76.0	85.6	90.0	56.2	64.5	72.5
	Pantheon Exterior	1401	76.1	76.8	77.3	79.2	85.6	88.7	60.8	66.5	72.2
	Reichstag	75	92.1	92.2	92.2	82.0	94.2	97.0	61.0	72.1	83.5
	Sacre Coeur	1179	32.1	38.3	44.3	51.3	63.4	75.8	24.7	27.2	32.4
	Saint Peter's Square	2504	73.5	79.0	79.5	42.6	71.0	84.4	18.0	33.8	55.5
	Taj Mahal	1312	62.0	64.4	65.5	65.0	80.8	88.9	44.2	49.0	56.0
	Temple Nara Japan	904	77.2	77.9	78.3	64.1	78.3	85.4	41.3	52.6	65.1
	Trevi Fountain	3191	56.0	57.3	59.0	61.5	71.1	74.3	37.4	43.9	50.8
	Average	-	68.6	70.8	72.7	66.2	78.6	85.0	43.5	52.1	61.6
Test	British Museum	660	99.4	99.7	100.0	63.1	87.7	94.7	43.2	59.1	80.0
	Florence Cathedral Side	108	99.7	100.0	100.0	82.2	95.5	98.4	60.3	74.8	89.3
	Lincoln Memorial Statue	850	100.0	100.0	100.0	85.6	96.4	98.6	58.2	75.1	89.9
	Milan Cathedral	124	100.0	100.0	100.0	73.4	94.0	98.2	50.8	67.7	86.4
	Mount Rushmore	138	99.9	100.0	100.0	41.5	71.3	86.5	27.8	40.5	65.1
	Piazza San Marco	249	99.9	100.0	100.0	87.9	97.9	99.3	68.1	81.0	92.6
	Sagrada Familia	401	99.5	100.0	100.0	70.6	91.5	96.6	48.6	65.1	84.0
	St. Paul's Cathedral	615	99.6	100.0	100.0	71.9	91.1	96.5	49.2	65.3	84.0
	Average	-	99.7	100.0	100.0	72.0	90.7	96.1	50.8	66.1	83.9

Table 19. **AMB3R-SfM results on the IMC Phototourism dataset [50].** Evaluation is performed on all available images per scene. Despite the unconstrained nature of photo-tourism collections, AMB3R-SfM shows robust generalization in a purely feed-forward manner, without relying on any test-time optimization.

9. Additional Quantitative Results

Ablation study. We present additional ablation studies in Tab. 17 to analyze the impact of key design choices, including the backend (2D vs. 3D), scale alignment for supervision, and zero convolution. Notably, training without zero convolution fails to converge under our current computational budget and dataset size. As discussed in Sec. 3, this is caused by catastrophic forgetting of the learned confidence: without zero convolution, the confidence function with randomly initialized weights of the backend will shift drastically, and lead to inconsistent learning objectives. In this case, convergence would likely require training resources and data comparable to those used for VGGT [126].

We also report additional ablations on the metric-scale

Approach	GT Poses	GT Range	GT Intrinsic	Align	KITTI rel ↓	KITTI $\delta_{1.03}$ ↑	ScanNet rel ↓	ScanNet $\delta_{1.03}$ ↑	ETH3D rel ↓	ETH3D $\delta_{1.03}$ ↑	DTU rel ↓	DTU $\delta_{1.03}$ ↑	T&T rel ↓	T&T $\delta_{1.03}$ ↑	Average rel ↓	Average $\delta_{1.03}$ ↑
a) Classic approaches																
COLMAP [101]	✗	✗	✗	✗	12.0	58.2	14.6	34.2	16.4	55.1	**0.7**	**96.5**	2.7	**95.0**	9.3	67.8
COLMAP Dense [101]	✗	✗	✗	✗	26.9	52.7	38.0	22.5	89.8	23.2	20.8	69.3	25.7	76.4	40.2	48.8
b) Single-view depth																
Depth Pro [6]	✗	✗	✓	med	6.1	39.6	(4.3)	(58.4)	6.1	53.5	5.6	49.6	5.6	57.5	5.6	51.7
Metric3D [41]	✗	✗	✓	med	5.1	44.1	2.4	78.3	4.4	54.5	10.1	39.5	6.2	48.0	5.6	52.9
UniDepthV2 [94]	✗	✗	✓	med	4.0	55.3	(2.1)	(82.6)	3.7	66.2	3.2	72.3	3.6	68.4	3.3	68.9
DepthAnything V2 [139]	✗	✗	✗	lstsq †	6.6	38.6	4.0	58.6	4.7	56.5	2.6	74.7	4.5	57.5	4.8	54.1
c) Depth from frames and poses (w/ per-image range)																
MVSNet [140]	✓	✓	✓	✗	22.7	36.1	24.6	20.4	35.4	31.4	(1.8)	(86.0)	8.3	73.0	18.6	49.4
Vis-MVSNet [147]	✓	✓	✓	✗	9.5	55.4	8.9	33.5	10.8	43.3	(1.8)	(87.4)	4.1	87.2	7.0	61.4
PatchmatchNet [123]	✓	✓	✓	✗	10.8	45.8	8.5	35.3	19.1	34.8	(2.1)	(82.8)	4.8	82.9	9.1	56.3
MVSFormer++ [15]	✓	✓	✓	✗	4.4	65.7	7.9	39.4	7.8	50.4	(0.9)	(95.3)	3.2	88.1	4.8	67.8
d) Depth from frames and poses (w/o per-image range)																
Fast-MVSNet [143]	✓	✗	✓	✗	12.1	37.4	287.1	9.4	131.2	9.6	(540.4)	(1.9)	33.9	47.2	200.9	21.1
Robust MVD Baseline [104]	✓	✗	✓	✗	7.1	41.9	7.4	38.4	9.0	42.6	2.7	82.0	5.0	75.1	6.3	56.0
MVSA [46]	✓	✗	✓	✗	3.2	68.8	3.7	62.9	3.2	68.0	1.3	95.0	2.1	90.5	2.7	77.0
e) Depth from frames (w/o poses)																
DeMoN [120]	✗	✗	✓	$\|\mathbf{t}\|$	15.5	15.2	12.0	21.0	17.4	15.4	21.8	16.6	13.0	23.2	16.0	18.3
DUSt3R [130]	✗	✗	✗	med	5.4	49.5	(3.1)	(71.8)	3.0	76.0	3.9	68.6	3.3	75.1	3.7	68.2
Spann3R [124]	✗	✗	✗	med	7.9	36.2	(3.3)	(67.1)	5.7	58.6	3.5	65.2	4.7	58.5	5.0	57.1
Pow3R [48]	✗	✗	✗	med	5.7	45.7	(3.2)	(68.8)	3.0	74.7	3.0	74.3	3.3	76.6	3.6	68.0
MUSt3R [13]	✗	✗	✗	med	4.5	55.0	(4.0)	(59.8)	2.5	80.3	4.6	55.4	(2.6)	(80.4)	3.7	66.2
VGGT [126]	✗	✗	✗	med	4.5	59.6	(2.3)	(80.8)	1.8	86.3	0.9	95.6	2.4	84.1	2.4	81.3
$\pi3^{‡}$ [131]	✗	✗	✗	med	2.8	72.9	(2.0)	(83.6)	**1.3**	**92.4**	1.3	91.8	1.8	87.3	1.8	85.6
MapAnything‡ [126]	✗	✗	✗	med	4.0	59.4	4.0	60.5	2.8	73.2	3.9	63.7	3.3	73.0	3.6	66.0
AMB3R	✗	✗	✗	med	**2.8**	**74.4**	**(1.9)**	**(85.8)**	1.4	90.9	0.9	95.1	**1.7**	90.2	**1.7**	**87.3**

Table 20. **Multi-view depth estimation.** Our method achieves state-of-the-art performance on RMVDB [104]. ‡ means concurrent works, and (parentheses) indicate. We only report a subset of a)-e) methods, and please refer to supplementary material for the full table.

	Scenes	COLMAP [101]		ACE-Zero [10]		FlowMap [108]		VGGSfM [126]		DF-SfM [39]		MASt3R-SfM [28]		AMB3R (SfM)	
		RRA@5	RTA@5	RRA@5	RTA@5	RRA@5	RTA@5	RRA@5	RTA@5	RRA@5	RTA@5	RRA@5	RTA@5	RRA@5	RTA@5
Training	Barn	GT	GT	56.1	55.6	-	-	-	-	100.	99.8	52.6	85.6	91.8	96.4
	Caterpillar	GT	GT	87.3	95.6	-	-	-	-	-	-	84.2	92.3	100.0	98.6
	Church	GT	GT	90.5	76.3	-	-	-	-	-	-	11.6	16.8	93.0	92.9
	Courthouse	GT	GT	44.1	45.0	-	-	-	-	-	-	8.8	9.9	79.9	77.7
	Ignatius	GT	GT	100.	99.9	62.5	70.0	-	-	100.	99.9	43.6	60.1	100.0	99.5
	Meetingroom	GT	GT	39.3	38.5	26.3	39.8	-	-	84.1	89.0	92.6	89.9	100.0	96.3
	Truck	GT	GT	100.	99.7	53.4	69.6	-	-	100.	99.8	100.	99.7	100.0	99.7
Intermediate	Family	GT	GT	38.9	44.6	-	-	-	-	-	-	22.3	25.9	100.0	99.6
	Francis	GT	GT	57.4	79.0	57.6	67.7	-	-	100.	99.7	17.0	41.0	100.0	92.1
	Horse	GT	GT	68.2	81.8	-	-	-	-	-	-	6.4	6.3	100.0	98.8
	Lighthouse	GT	GT	30.6	38.8	4.8	9.5	-	-	66.3	66.0	50.8	72.1	100.0	98.7
	M60	GT	GT	100.	99.9	50.4	48.3	-	-	100.	99.8	100.	100.	100.0	99.0
	Panther	GT	GT	100.	99.5	100.	77.6	-	-	100.	99.1	100.	99.5	100.0	99.6
	Playground	GT	GT	82.7	85.5	49.1	63.8	-	-	100.	99.9	99.3	99.3	100.0	98.3
	Train	GT	GT	62.6	62.5	18.4	29.2	-	-	42.8	41.8	10.6	15.8	89.4	88.8
Advanced	Auditorium	GT	GT	1.6	1.1	1.3	1.4	-	-	1.7	1.7	1.7	1.5	96.0	88.7
	Ballroom	GT	GT	56.4	43.2	14.1	16.7	-	-	56.0	44.4	43.8	29.6	35.2	37.5
	Courtroom	GT	GT	62.5	54.1	5.3	3.6	-	-	66.8	66.3	67.2	69.1	38.7	51.1
	Museum	GT	GT	13.5	11.1	0.8	1.2	-	-	14.8	13.5	12.3	11.0	100.0	98.4
	Palace	GT	GT	3.1	3.9	-	-	-	-	25.6	27.7	27.0	35.7	38.6	61.4
	Temple	GT	GT	0.4	0.9	0.5	1.2	-	-	55.5	60.7	80.7	72.2	99.3	96.9

Table 21. **Detailed per-scene SfM results on Tanks and Temples [57].** We use COLMAP GT [99] for evaluation as in MASt3R-SfM [28]. The baseline results are obtained from MASt3R-SfM. (-) indicates failure.

head in Tab. 18, in addition to the scale-difference vs. median-z comparison presented in Tab. 4.

Multi-view depth estimation. We show the full table of multi-view depth estimation results in Tab. 20 with detailed

input modality and the alignment.

Structure from motion. We show per-scene decomposition of SfM on Tanks&Temples [57] and IMC Phototourism [50] in Tab. 21 and Tab. 19. The COLMAP results [99] are used as the ground-truth.

10. Qualitative examples

Visual Odometry. As in Fig. 9 and Fig. 10, we present qualitative results of our visual odometry on all static scenes from the 7Scenes [107], TUM [110], and ETH SLAM [103] datasets used in our evaluation.

Structure from Motion. Fig. 11 and Fig. 12 show qualitative results of SfM on all scenes from the ETH3D [102] and Tanks&Temples [57] datasets. In Fig. 13, we further include qualitative results on the Cambridge Landmarks dataset, in-the-wild image collections, and the IMC Phototourism [50] dataset, which contains large-scale unordered images captured by tourists.

Figure 9. **Qualitative showcase** of VO on 7scenes [107]. We visualize keyframe poses as red cameras and non-keyframe poses as green dots.

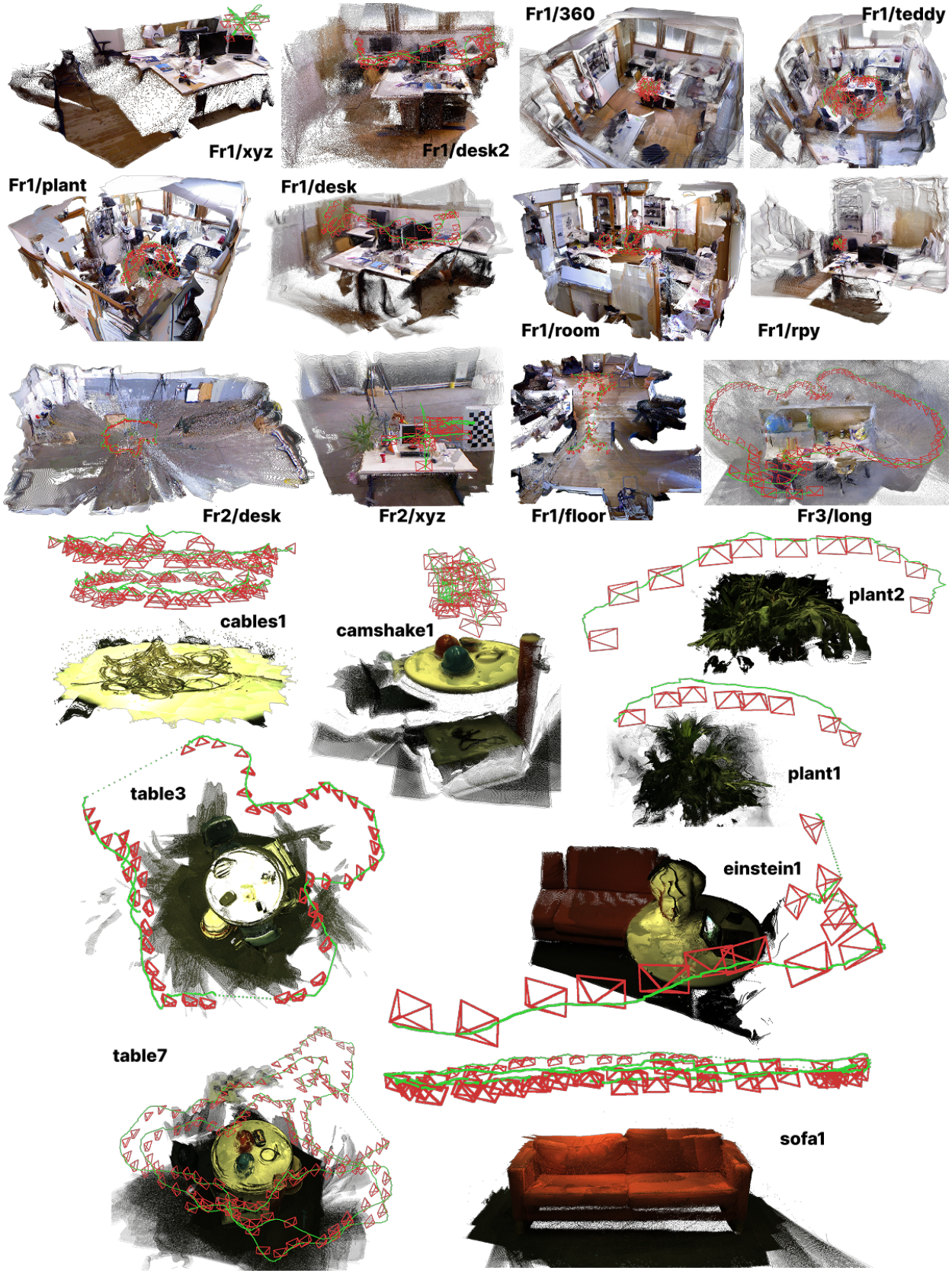

Figure 10. **Qualitative showcase** of VO on TUM [110] and ETH SLAM [103] datasets.

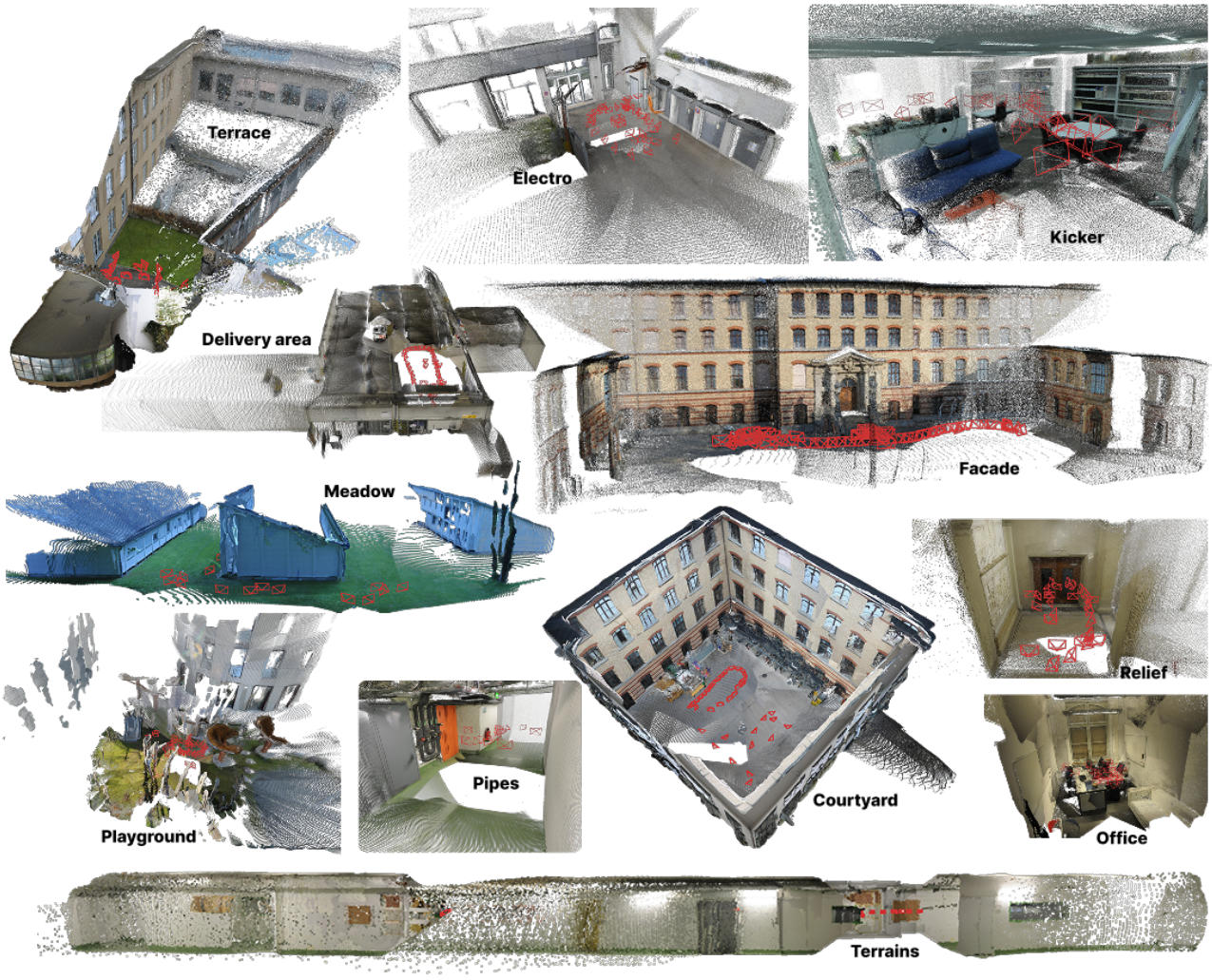

Figure 11. **Qualitative showcase** of structure from motion on ETH3D [102]

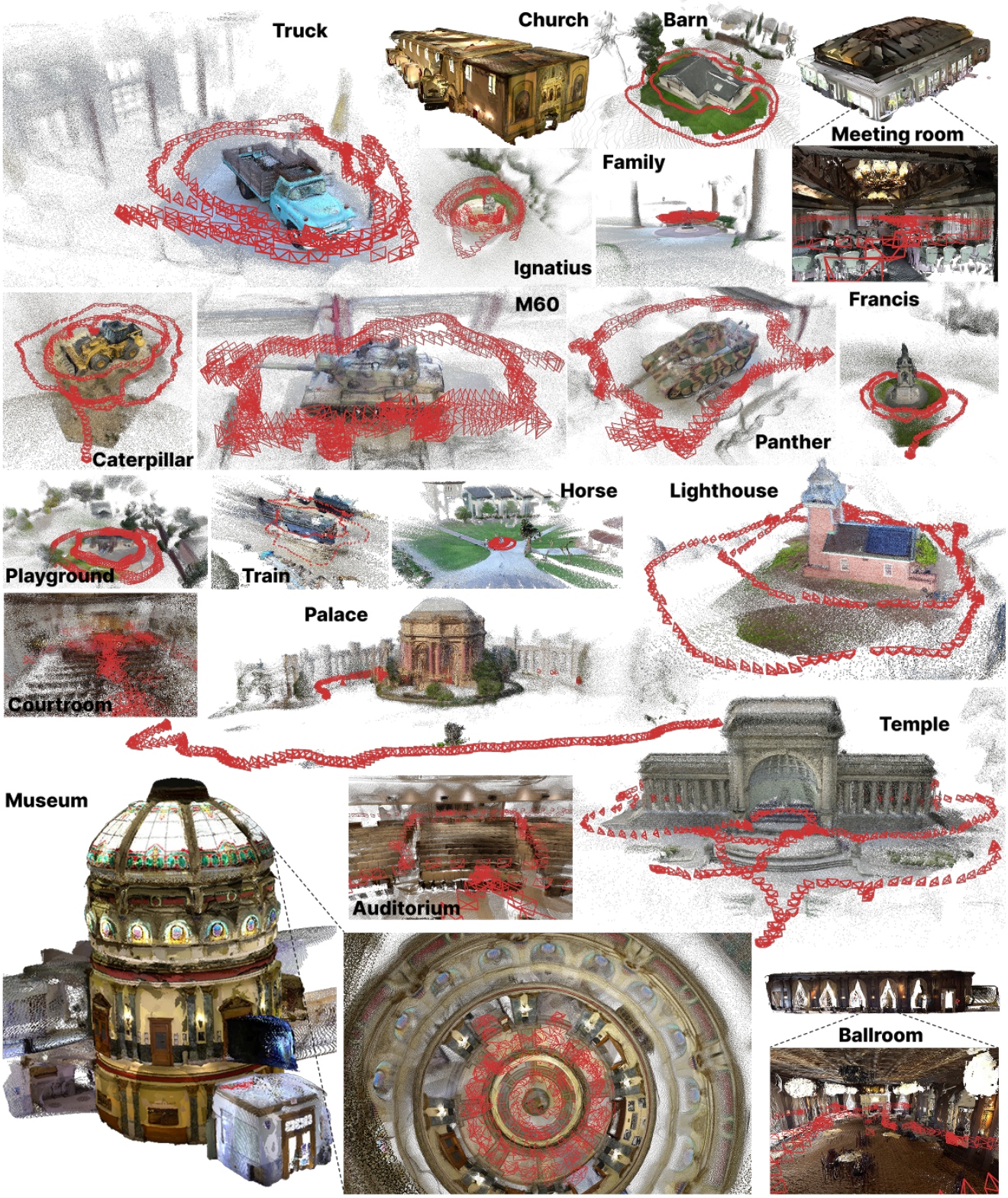

Figure 12. **Qualitative showcase** of structure from motion on Tanks and Temples [57] dataset.

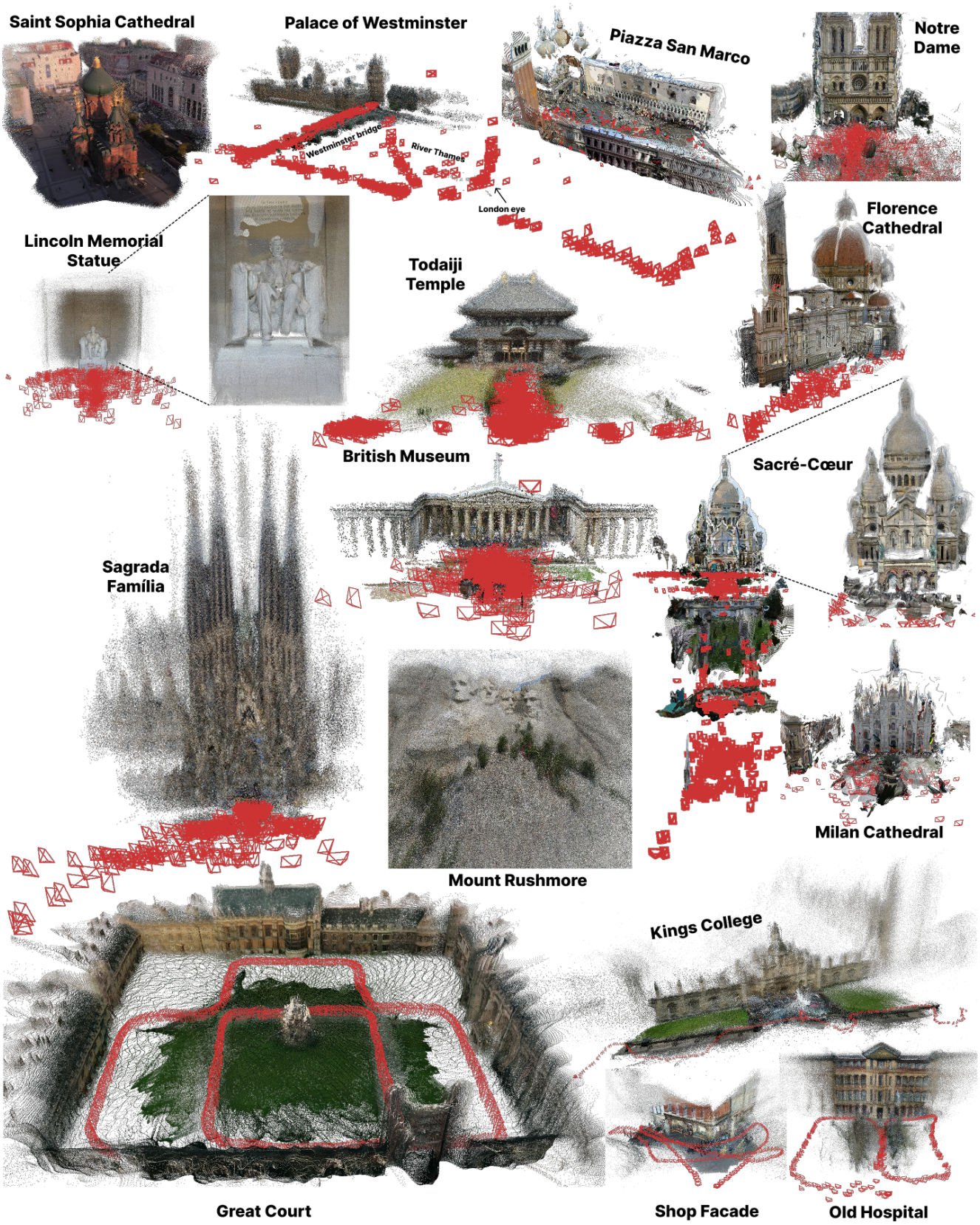

Figure 13. **Qualitative showcase** of structure from motion on IMC phototourism [50], Cambridge landmarks [55]. The results are randomly downsampled to 3 million points for visualization purpose.

References

[1] 3Blue1Brown. Hilbert's curve: Is infinite math useful?, 2017. Accessed: 2025-09-26. 2

[2] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016. 5, 6, 12, 13

[3] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M Seitz, and Richard Szeliski. Building rome in a day. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 72–79, 2009. 1, 2

[4] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 11

[5] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 5

[6] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 12, 14

[7] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6684–6692, 2017. 2

[8] Eric Brachmann, Martin Humenberger, Carsten Rother, and Torsten Sattler. On the limits of pseudo ground truth in visual camera re-localisation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 6218–6228, 2021. 7, 12

[9] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5044–5053, 2023. 2

[10] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. *arXiv preprint arXiv:2404.14351*, 2024. 2, 8, 12, 13, 14

[11] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 611–625. Springer, 2012. 6, 13

[12] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 11

[13] Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1050–1060, 2025. 2, 6, 7, 8, 12, 13, 14

[14] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34–44, 2023. 10

[15] Chenjie Cao, Xinlin Ren, and Yanwei Fu. Mvsformer++: Revealing the devil in transformer's details for multi-view stereo. *arXiv preprint arXiv:2401.11673*, 2024. 6, 12, 14

[16] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, 2022. 2

[17] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 2

[18] Wanli Chen, Xinge Zhu, Guojin Chen, and Bei Yu. Efficient point cloud analysis using hilbert curve. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 730–747. Springer, 2022. 2

[19] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 1, 2

[20] Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. *arXiv preprint arXiv:2507.18255*, 2025. 2

[21] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5 (2):721–728, 2020. 7, 13

[22] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5828–5839, 2017. 5, 6, 11, 12

[23] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067, 2007. 1, 2

[24] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. Sail-recon: Large sfm by augmenting scene regression with localization. *arXiv preprint arXiv:2508.17972*, 2025. 2

[25] Eric Dexheimer and Andrew J Davison. Learning a depth covariance function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13122–13131, 2023. 7, 13

[26] Eric Dexheimer and Andrew J Davison. Como: Compact mapping and odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 349–365. Springer, 2024. 7, 13

[27] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16739–16752, 2025. 2

[28] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025. 2, 8, 13, 14

[29] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10786–10796, 2021. 5, 12

[30] Sven Elflein, Qunjie Zhou, and Laura Leal-Taixé. Light3r-sfm: Towards feed-forward structure-from-motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16774–16784, 2025. 2

[31] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 40(3):611–625, 2017. 7, 13

[32] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 2

[33] Xianze Fang, Jingnan Gao, Zhe Wang, Zhuo Chen, Xingyu Ren, Jiangjing Lyu, Qiaomu Ren, Zhonglei Yang, Xiaokang Yang, Yichao Yan, et al. Dens3r: A foundation model for 3d geometry prediction. *arXiv preprint arXiv:2507.16290*, 2025. 2

[34] Xin Fei, Wenzhao Zheng, Yueqi Duan, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Jiwen Lu. Driv3r: Learning dense 4d reconstruction for autonomous driving. *arXiv preprint arXiv:2412.06777*, 2024. 2

[35] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 1

[36] Gonzalo Martin Garcia, Karim Abou Zeid, Christian Schmidt, Daan De Geus, Alexander Hermans, and Bastian Leibe. Fine-tuning image-conditional diffusion models is easier than you think. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 753–762. IEEE, 2025. 5, 12

[37] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. 1, 5, 6, 12, 13

[38] Michael Grupp. evo: Python package for the evaluation of odometry and slam., 2017. 6

[39] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21594–21603, 2024. 8, 13, 14

[40] David Hilbert. Über die stetige abbildung einer linie auf ein flächenstück. In *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes: Nebst Einer Lebensgeschichte*, pages 1–2. Springer, 1935. 2

[41] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2024. 14

[42] Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *arXiv preprint arXiv:2505.22657*, 2025. 2

[43] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. 2

[44] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2821–2830, 2018. 11

[45] Rui Huang, Guangyao Zhai, Zuria Bauer, Marc Pollefeys, Federico Tombari, Leonidas Guibas, Gao Huang, and Francis Engelmann. Video perception models for 3d scene synthesis. *arXiv preprint arXiv:2506.20601*, 2025. 2

[46] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11493–11504, 2025. 6, 12, 14

[47] Chris L Jackins and Steven L Tanimoto. Oct-trees and their use in representing three-dimensional objects. *Computer Graphics and Image Processing*, 14(3):249–270, 1980. 2

[48] Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1071–1081, 2025. 2, 6, 12, 14

[49] Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025. 2

[50] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Im-

age matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2): 517–547, 2021. 1, 13, 15, 20

[51] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Foundations of Crystallography*, 32(5): 922–923, 1976. 4

[52] Ben Kaye, Tomas Jakab, Shangzhe Wu, Christian Ruprecht, and Andrea Vedaldi. Dualpm: dual posed-canonical point maps for 3d shape and pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6425–6435, 2025. 2

[53] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 5, 12

[54] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2, 6, 12, 13

[55] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015. 20

[56] Ramil Khafizov, Artem Komarichev, Ruslan Rakhimov, Peter Wonka, and Evgeny Burnaev. G-cut3r: Guided 3d reconstruction with camera and depth prior integration. *arXiv preprint arXiv:2508.11379*, 2025. 2

[57] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36 (4), 2017. 1, 5, 6, 8, 12, 13, 14, 15, 19

[58] Lukas Koestler, Nan Yang, Niclas Zeller, and Daniel Cremers. Tandem: Tracking and dense mapping in real-time using deep multi-view stereo. In *Conference on Robot Learning*, pages 34–45. PMLR, 2022. 7, 13

[59] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025. 2

[60] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 71–91. Springer, 2024. 2, 5, 6, 13

[61] Haodong Li, Chen Wang, Jiahui Lei, Kostas Daniilidis, and Lingjie Liu. Stereodiff: Stereo-diffusion synergy for video depth estimation. *arXiv preprint arXiv:2506.20756*, 2025. 2

[62] Rui Li, Biao Zhang, Zhenyu Li, Federico Tombari, and Peter Wonka. Lari: Layered ray intersections for single-view 3d geometric reasoning. *arXiv preprint arXiv:2504.18424*, 2025. 2

[63] Zhiqi Li, Chengrui Dong, Yiming Chen, Zhangchi Huang, and Peidong Liu. Vicasplat: A single run is all you need for 3d gaussian splatting and camera estimation from unposed video frames. *arXiv preprint arXiv:2503.10286*, 2025. 2

[64] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10486–10496, 2025. 8, 13

[65] Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool. *arXiv preprint arXiv:2509.05296*, 2025. 2

[66] Zhen Colin Li, Weiwei Sun, Shrisudhan Govindarajan, Shaobo Xia, Daniel Rebain, Kwang Moo Yi, and Andrea Tagliasacchi. Noksr: Kernel-free neural surface reconstruction via point cloud serialization. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025. 2

[67] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065*, 2025. 2

[68] Chin-Yang Lin, Cheng Sun, Fu-En Yang, Min-Hung Chen, Yen-Yu Lin, and Yu-Lun Liu. Longsplat: Robust unposed 3d gaussian splatting for casual long videos. *arXiv preprint arXiv:2508.14041*, 2025. 2

[69] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 424–440. Springer, 2024. 7, 13

[70] Sidun Liu, Wenyu Li, Peng Qiao, and Yong Dou. Regist3r: Incremental registration with stereo foundation model. *arXiv preprint arXiv:2504.12356*, 2025. 2

[71] Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16651–16662, 2025. 2

[72] Thibaut Loiseau, Guillaume Bourmaud, and Vincent Lepetit. Alligat0r: Pre-training through co-visibility segmentation for relative camera pose regression. *arXiv preprint arXiv:2503.07561*, 2025. 2

[73] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22820–22830, 2025. 2

[74] Ziqi Lu, Heng Yang, Danfei Xu, Boyi Li, Boris Ivanovic, Marco Pavone, and Yue Wang. Lora3d: Low-rank self-calibration of 3d geometric foundation models. *arXiv preprint arXiv:2412.07746*, 2024. 2

[75] Jiahao Ma, Lei Wang, David Ahmedt-Aristizabal, Chuong Nguyen, et al. Puzzles: Unbounded video-depth augmen-

tation for scalable end-to-end 3d reconstruction. *arXiv preprint arXiv:2506.23863*, 2025. 2

[76] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 4, 7, 8

[77] Soroush Mahdi, Fardin Ayar, Ehsan Javanmardi, Manabu Tsukada, and Mahdi Javanmardi. Evict3r: Training-free token eviction for memory-bounded streaming visual geometry transformers. *arXiv preprint arXiv:2509.17650*, 2025. 2

[78] Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025. 2

[79] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. Gaussian splatting slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18039–18048, 2024. 7, 13

[80] Donald Meagher. Geometric modeling using octree encoding. *Computer graphics and image processing*, 19(2):129–147, 1982. 2

[81] Andreas Meuleman, Ishaan Shah, Alexandre Lanvin, Bernhard Kerbl, and George Drettakis. On-the-fly reconstruction for large-scale novel view synthesis from unposed images. *ACM Transactions on Graphics (TOG)*, 44(4):1–14, 2025. 2

[82] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421, 2020. 1

[83] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 2

[84] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 7, 13

[85] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 2, 7, 8, 13

[86] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 5, 6, 12

[87] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 1, 7, 12

[88] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013. 2

[89] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7855–7862. IEEE, 2019. 6, 13

[90] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[91] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 11

[92] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 1

[93] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 523–540. Springer, 2020. 2

[94] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 6, 12, 14

[95] Quanhao Qian, Guoyang Zhao, Gongjie Zhang, Jiuniu Wang, Ran Xu, Junlong Gao, and Deli Zhao. Gp3: A 3d geometry-aware policy with multi-view images for robotic manipulation. *arXiv preprint arXiv:2509.15733*, 2025. 2

[96] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 11

[97] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 10901–10911, 2021. 11

[98] Jerome Revaud, Yohann Cabon, Romain Brégier, JongMin Lee, and Philippe Weinzaepfel. Sacreg: Scene-agnostic coordinate regression for visual localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 688–698, 2024. 2

[99] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 8, 14, 15

[100] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic

synthetic dataset for holistic indoor scene understanding. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 11

[101] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. 1, 2, 6, 8, 10, 12, 13, 14

[102] Thomas Schops, Johannes L Schonberger, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3260–3269, 2017. 5, 6, 8, 12, 13, 15, 18

[103] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 134–144, 2019. 6, 8, 13, 15, 17

[104] Philipp Schröppel, Jan Bechtold, Artemij Amiranashvili, and Thomas Brox. A benchmark and a baseline for robust multi-view depth estimation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 637–645. IEEE, 2022. 5, 6, 12, 13, 14

[105] Duochao Shi, Weijie Wang, Donny Y Chen, Zeyu Zhang, Jia-Wang Bian, Bohan Zhuang, and Chunhua Shen. Revisiting depth representations for feed-forward 3d gaussian splatting. *arXiv preprint arXiv:2506.05327*, 2025. 2

[106] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, 2013. 6

[107] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2937, 2013. 2, 5, 12, 13, 15, 16

[108] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024. 13, 14

[109] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[110] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, 2012. 6, 8, 11, 13, 15, 17

[111] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15598–15607, 2021. 1

[112] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020. 11

[113] Yang-Tian Sun, Xin Yu, Zehuan Huang, Yi-Hua Huang, Yuan-Chen Guo, Ziyi Yang, Yan-Pei Cao, and Xiaojuan Qi. Unigeo: Taming video diffusion for unified consistent geometry estimation. *arXiv preprint arXiv:2505.24521*, 2025. 2

[114] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 12

[115] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5283–5293, 2025. 2

[116] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 2

[117] Zachary Teed and Jia Deng. Deepv2d: Video to depth with differentiable structure from motion. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 7, 13

[118] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:16558–16569, 2021. 7, 13

[119] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(4):376–380, 2002. 4

[120] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5038–5047, 2017. 6, 12, 14

[121] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019. 5, 12

[122] Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Faster vggt with block-sparse global attention. *arXiv preprint arXiv:2509.07120*, 2025. 2

[123] Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. Patchmatchnet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14194–14203, 2021. 12, 14

[124] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 78–89. IEEE, 2025. 1, 2, 3, 5, 6, 7, 8, 12, 13, 14

[125] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13293–13302, 2023. 1, 6, 13

[126] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306, 2025. 1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14

[127] Kaixuan Wang and Shaojie Shen. Flow-motion and depth network for monocular stereo and beyond. *IEEE Robotics and Automation Letters*, 5(2):3307–3314, 2020. 11

[128] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10510–10522, 2025. 2, 6, 7, 13

[129] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5261–5271, 2025. 3, 4, 5, 12

[130] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 1, 2, 3, 5, 6, 12, 14

[131] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 2, 6, 7, 12, 13, 14

[132] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 803–814, 2023. 11

[133] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:33330–33342, 2022. 2

[134] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4840–4851, 2024. 2, 4

[135] Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *arXiv preprint arXiv:2507.02863*, 2025. 2

[136] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22378–22389, 2024. 11

[137] Jiale Xu, Shenghua Gao, and Ying Shan. Freesplatter: Pose-free gaussian splatting for sparse-view 3d reconstruction. *arXiv preprint arXiv:2412.09573*, 2024. 2

[138] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21924–21935, 2025. 2, 5

[139] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:21875–21911, 2024. 5, 6, 12, 14

[140] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 12, 14

[141] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 11

[142] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 12

[143] Zehao Yu and Shenghua Gao. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1949–1958, 2020. 12, 14

[144] Yijun Yuan, Zhuoguang Chen, Kenan Li, Weibang Wang, and Hang Zhao. Slam-former: Putting slam into one transformer. *arXiv preprint arXiv:2509.16909*, 2025. 2

[145] Ganlin Zhang, Erik Sandström, Youmin Zhang, Manthan Patel, Luc Van Gool, and Martin R Oswald. Glorie-slam: Globally optimized rgb-only implicit encoding point cloud slam. *arXiv preprint arXiv:2403.19549*, 2024. 7, 13

[146] Ganlin Zhang, Shenhan Qian, Xi Wang, and Daniel Cremers. Vista-slam: Visual slam with symmetric two-view association. *arXiv preprint arXiv:2509.01584*, 2025. 2

[147] Jingyang Zhang, Shiwei Li, Zixin Luo, Tian Fang, and Yao Yao. Vis-mvsnet: Visibility-aware multi-view stereo network. *International Journal of Computer Vision (IJCV)*, 131(1):199–214, 2023. 6, 12, 14

[148] Jiakai Zhang, Shouchen Zhou, Haizhao Dai, Xinhang Liu, Peihao Wang, Zhiwen Fan, Yuan Pei, and Jingyi Yu. Cryofastar: Fast cryo-em ab initio reconstruction made easy. *arXiv preprint arXiv:2506.05864*, 2025. 2

[149] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 2

[150] Songyan Zhang, Yongtao Ge, Jinyuan Tian, Guangkai Xu, Hao Chen, Chen Lv, and Chunhua Shen. Pomato: Marrying pointmap matching with temporal motion for dynamic 3d reconstruction. *arXiv preprint arXiv:2504.05692*, 2025. 2

[151] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21936–21947, 2025. 2

[152] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3727–3737, 2023. 7, 13

[153] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 2

[154] Xinyi Zheng, Steve Zhang, Weizhe Lin, Aaron Zhang, Walterio W Mayol-Cuevas, Yunze Liu, and Junxiao Shen. Culture3d: A large-scale and diverse dataset of cultural landmarks and terrains for gaussian-based scene rendering. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 29064–29074, 2025. 7

[155] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 5

[156] Yuxuan Zhou, Xingxing Li, Shengyu Li, Zhuohao Yan, Chunxi Xia, and Shaoquan Feng. Mast3r-fusion: Integrating feed-forward visual model with imu, gnss for high-functionality slam. *arXiv preprint arXiv:2509.20757*, 2025. 2

[157] Haodong Zhu, Changbai Li, Yangyang Ren, Zichao Feng, Xuhui Liu, Hanlin Chen, Xiantong Zhen, and Baochang Zhang. Surf3r: Rapid surface reconstruction from sparse rgb views in seconds. *arXiv preprint arXiv:2508.04508*, 2025. 2

[158] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12786–12796, 2022. 6

[159] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 2

[160] Lojze Zust, Yohann Cabon, Juliette Marrie, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud, and Gabriela Csurka. Panst3r: Multi-view consistent panoptic segmentation. *arXiv preprint arXiv:2506.21348*, 2025. 2