Long-Tail Internet Photo Reconstruction

Yuan Li[1] Yuanbo Xiangli[1†] Hadar Averbuch-Elor[1] Noah Snavely[1] Ruojin Cai[2†]
[1]Cornell University [2]Kempner Institute, Harvard University

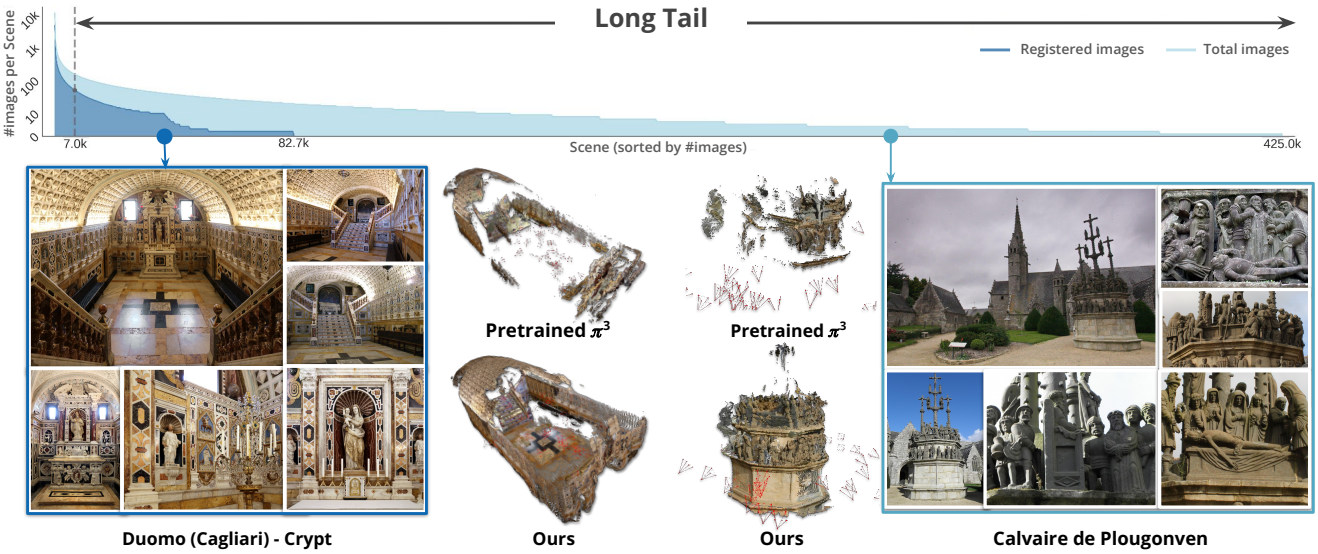

Figure 1. **Long-tail Internet photo reconstruction.** Internet photo collections follow a long-tailed distribution. In the top plot, the x-axis represents scene index (sorted by image count) and the y-axis shows images per scene (scenes are drawn from MegaScenes [27], a dataset of Internet photo collections). The light blue curve plots the total number of Internet photos per scene, while the steel blue curve shows the size of the subset of photos that were successfully registered using SfM. The *head* of this distribution of photo collections represents well-photographed scenes; here, there are 6,985 scenes with $>$50 registered images. However, most photo collections are in the *long tail* of this distribution; here, 418,056 scenes with fewer than 50 registered photos. State-of-the-art methods often fail on scenes in this tail. In the lower half of the figure, we show two examples from the long tail, along with representative input images and the corresponding reconstructions. On Calvaire de Plougonven, COLMAP doesn't register any image; on both Duomo (Cagliari)-Crypt and Calvaire de Plougonven, recent feed-forward reconstruction models like π^3 [34] produce poor results. We propose MegaDepth-X dataset and a strategy for mimicking long-tail camera distributions, on which fine-tuned models like π^3 exhibit better reconstruction robustness.

Abstract

Internet photo collections exhibit an extremely long-tailed distribution: a few famous landmarks are densely photographed and easily reconstructed in 3D, while most real-world sites are represented with sparse, noisy, uneven imagery beyond the capabilities of both classical and learned 3D methods. We believe that tackling this long-tail regime represents one of the next frontiers for 3D foundation models. Although reliable ground-truth 3D supervision from sparse scenes is challenging to acquire, we observe that it can be effectively simulated by sampling sparse subsets from well-reconstructed Internet landmarks. To this end, we introduce MegaDepth-X, a large dataset of 3D reconstructions with clean, dense depth, together with a strategy for sampling sets of training images that mimic camera distributions in long-tail scenes. Finetuning 3D foundation models with these components yields robust reconstructions under extreme sparsity, and also enables more reliable reconstruction in symmetric and repetitive scenes, while preserving generalization to standard, dense 3D benchmark datasets. The dataset, finetuned models, and code are available at: `megadepth-x.github.io`.

1. Introduction

Internet photo collections of real-world landmarks follow a long-tailed distribution. A small fraction of famous sites, such as the Colosseum or Notre Dame, are photographed

†Corresponding authors.

from every conceivable angle and can be accurately reconstructed by standard Structure-from-Motion (SfM) pipelines. Yet the overwhelming majority of landmarks across the world are represented on the Internet with just a handful of sparse, noisy images (Fig. 1). We refer to this large body of scenes as the *long-tail* of online photo collections. Such scenes are the norm rather than the exception in real-world Internet imagery.

Reconstructing long-tail scenes is challenging. Classic methods, such as COLMAP [19], often fail because feature correspondence is hard to find across sparse, non-overlapping, or wide-baseline views. Modern learned feed-forward models, like DUSt3R [33] and VGGT [30], can learn powerful priors from millions of images that might help reconstruct long-tail collections. In practice, however, these models are primarily trained on controlled captures with clean, dense, and evenly sampled data. When applied to long-tail Internet scenes featuring sparse, diverse, and unevenly distributed imagery, we find that these models often fail to recover consistent geometry.

We believe that one of the next frontiers for 3D foundation models lies in tackling this long-tail regime of Internet photos. Better data is almost certainly key to this problem, but we cannot easily construct reliable 3D supervision from long-tail collections themselves, as most contain too few overlapping views for robust reconstruction. Instead, we propose to *simulate* such long-tailed sets by appropriate sampling of sparse images from the large, well-reconstructed Internet landmarks at the head of the distribution, inheriting ground truth from the full reconstruction.

This strategy requires drawing from large amounts of high-quality landmark reconstructions from Internet photos. Existing datasets fall short of this need: MegaDepth [14] is clean but small, while MegaScenes [27] is large but noisy and lacks depth maps. We therefore introduce *MegaDepth-X* (dubbed MD-X), a next-generation extension of MegaDepth in both scale ($8\times$ larger) and quality: a large-scale, clean, and dense-depth-enhanced dataset built from Internet photo reconstructions with consistent depth refinement and extensive manual verification against reliable references (e.g., Google Maps and satellite imagery). Equipped with MD-X, we propose a novel *sparsity-aware* sampling strategy that mimics the camera distributions of long-tail scenes, encouraging training batches to span wide baselines and partial overlap rather than clustered dense views.

Through extensive experiments, we show that models fine-tuned with MD-X and our sparsity-aware data sampling scheme are significantly more robust on long-tail Internet photo collections, including challenging doppelganger scenes with ambiguous or symmetric content, such as the Calvaire de Plougonven example in Fig. 1, where classical SfM and pretrained foundation models often fail. In summary, our contributions are:

- **Defining the 3D long-tail regime**: we formalize and characterize the long-tail distribution of Internet photo collections, highlighting this setting's distinct challenges.
- **MegaDepth-X**, dubbed MD-X, a large-scale, clean, and depth-augmented dataset for finetuning 3D foundation models on real-world Internet scenes.
- **Sparsity-aware sampling** strategies that simulate the distribution of long-tail Internet collections to improve generalization of 3D prediction models on real-world data.

2. Related Work

Feed-forward 3D reconstruction. Reconstructing 3D scene geometry from 2D images is a cornerstone of computer vision. Traditional structure from motion (SfM) [20] and multi-view stereo (MVS) [21] methods were crowning achievements of the classic era of 3D vision, and were scaled to large Internet photo collections [1, 8, 25]. Recently, the new paradigm of feed-forward 3D reconstruction has emerged, which involves regressing 3D attributes directly from images in a single pass. Pioneering work in this area, such as DUSt3R, showed success at predicting pixel-aligned point maps from image pairs [33]. MASt3R extended this approach but still relied on pairwise processing [13]. Subsequent efforts focused on scaling these models to arbitrary numbers of views. VGGT [30], along with concurrent models like Fast3R [36] and FLARE [37], introduced large transformer architectures that can process hundreds of views simultaneously. By leveraging large-scale, diverse datasets and a multi-task learning objective, VGGT predicts a full suite of 3D attributes, including camera parameters, depth maps, and point maps. To eliminate reference-frame bias, π^3 [34] recently proposed a permutation-equivariant architecture that predicts affine-invariant camera poses and scale-invariant local point maps. These methods work well on small-scale, densely-captured, well-conditioned scenes. However, we find that their performance on more sparse and noisy Internet photos remains suboptimal, particularly for long-tail scenes.

Long-tail challenges in 3D vision. Long-tailed problems are pervasive in computer vision. They occur when data for common scenarios (the head) are abundant, but examples of rare yet collectively frequent cases (the tail) are scarce. For instance, many object recognition problems involve a few dominant categories but many rarely seen ones, and in autonomous driving, routine driving scenes are plentiful while safety-critical events are hard to capture.

Recently, MegaScenes [27] introduced a large-scale scene-level dataset built from Internet photo collections, where long-tail effects are particularly pronounced. Many scenes in the dataset are either unreconstructed or incorrectly reconstructed. These failures stem from a combination of view sparsity, noisy imagery, and doppelganger is-

sues [6]. Recent work has sought to address such challenges by developing stronger local features [7, 28] and matchers [10, 11, 15, 18], and by learning wide-baseline pose relationships from large-scale 3D datasets [3, 5]. The doppelganger problem was further addressed by Cai et al. [6, 35], who trained classifiers to prune false matches during the structure-from-motion phase of reconstruction.

While these advances have led to enhanced robustness, they do not yet work reliably at scale. Ideally, we'd mine ground truth 3D training data for long tail scenes and learn to reconstruct them, but that involves a chicken-and-egg problem, because the common practice of using available reconstructors (e.g. COLMAP [19, 22], VGGT [30]) to derive pseudo-ground-truth camera poses and point maps from natural data doesn't work. Instead, similar in spirit to approaches used in autonomous driving that augment training data by simulating rare events, our key idea is to take large, well-conditioned image collections and subsample them to simulate long-tailed photo collections, and use these to better balance training scene distributions for regression models in order to generalize to long-tailed scenes.

3. The MegaDepth-X Dataset

Learning in the long-tail regime requires high-quality 3D supervision derived from Internet photo collections. This involves two key challenges. First, reconstructions of Internet photo collections can be unreliable due to noise, dynamic content, and ambiguities [6]. Second, most long-tail scenes lack any usable reconstructions, as classical SfM pipelines like COLMAP [20] often fail on sparse or widely varying image sets. To address these issues, we construct MD-X, a large-scale, clean, and depth-refined dataset that provides reliable 3D supervision, built from well-reconstructed scenes in MegaScenes [27].

3.1. Filtering and Disambiguation

Our first step in constructing MD-X is to identify candidate Internet landmarks from which reliable supervision can be derived. We take as our starting pool the subset of MegaScenes with more than 100 registered images, which typically yields stable reconstructions. However, even these "well-reconstructed" scenes exhibit two common failure modes: (1) Many scenes contain dynamic events or crowded activities, causing feature matches to lock onto moving objects rather than static structures, leading to unreliable reconstructions. (2) The Doppelganger problem [6, 35], where visually similar but geographically distant images are mistakenly registered together. Both issues produce incorrect camera poses and fragmented, inconsistent point clouds as shown in Fig. 2.

To mitigate these issues, we first inspect the dataset and exclude scenes dominated by crowds or moving objects. Next, we address the doppelganger problem by replacing the default COLMAP SfM reconstruction with MASt3R-SfM [13], combined with Doppelganger classification [35]. Specifically, MASt3R-SfM constructs the scene graph using feature matches derived from MASt3R descriptors, after which the Doppelganger classifier identifies and prunes suspicious edges that may result from doppelganger-induced false correspondences. Finally, we manually verify the reconstructed scenes against external references such as Google Maps and satellite imagery, discarding any scenes that do not align with the corresponding bird's-eye view.

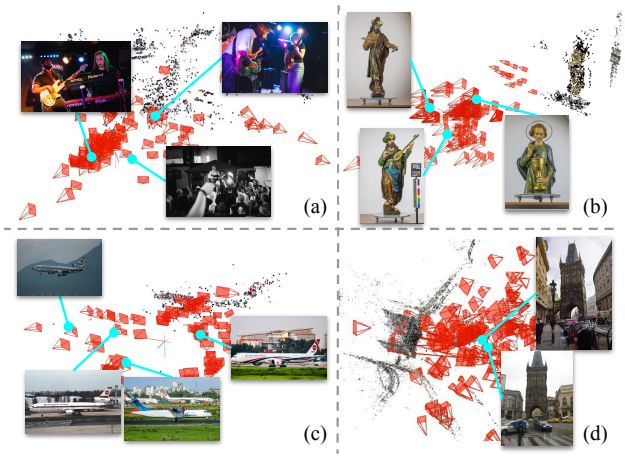

Figure 2. **Unreliable reconstructions in MegaScenes.** Reconstructions are unreliable when feature matches are incorrectly established on salient, non-static objects (e.g., (a) humans, (b) statues, (c) airplanes) instead of the static scene structure. This results in fragmented and geometrically inconsistent point clouds. Example (d) illustrates a doppelganger failure, where images from opposite sides of the building are incorrectly registered together.

3.2. Dense Depth Refinement

After obtaining reliable sparse reconstructions, we seek to generate dense depth maps for supervision. We start by running a standard multi-view stereo (MVS) [22] pipeline. We observe, as in prior work [14], that the resulting geometric depth maps from in-the-wild collections often exhibit artifacts, including depth-bleeding effects (background depths leak into foreground regions) and inconsistent and noisy depths in areas with transient objects (e.g., people, cars).

To address these initial issues, we apply the full depth refinement strategy from MegaDepth [14], including a modified MVS procedure that conservatively retains the minimum depth value during propagation, stability filtering to remove flickering pixels, and semantic filtering to exclude transient objects. However, even after this pipeline, we still observe artifacts in the processed geometric depth maps: (1) the MegaDepth-modified MVS still leads to depth-bleeding artifacts, and (2) semantic filtering is not ideal as it relies on a manually designated list of object categories. Examples of such issues are shown in Fig. 3.

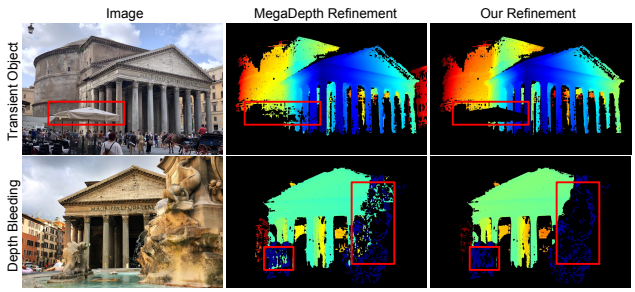

Figure 3. **Depth refinement.** MVS depth maps often suffer from artifacts like noise from transient objects (top row) and depth bleeding (bottom row). As shown in the middle column, the MegaDepth refinement pipeline (modified MVS, stability filtering, and semantic filtering) fails to fully remedy these issues. Our method (right column) introduces an additional monocular depth-guided filtering step, which effectively removes transient objects and significantly mitigates depth-bleeding artifacts.

Therefore, to augment MegaDepth's depth refinement procedure, we propose a monocular depth-guided filtering step. We use depth predictions from MoGe2 [32] as ordinal depth priors, and remove pixels in the processed geometric depth maps that are inconsistent with these priors. Specifically, we first align the processed geometric depths D_{geom} to the monocular predictions D_{mono} by matching their median values over valid pixels: $D'_{\text{geom}}(p) = s \cdot D_{\text{geom}}(p)$, where $s = \frac{\text{med}\{D_{\text{mono}}(p)|p\in P\}}{\text{med}\{D_{\text{geom}}(p)|p\in P\}}$. After scale alignment, we compute the normalized depth discrepancy between the two maps: $\Delta(p) = \frac{|D'_{\text{geom}}(p) - D_{\text{mono}}(p)|}{D'_{\text{geom}}(p)}$, and discard pixels whose discrepancies exceed a predefined threshold τ_{depth}. Moreover, to leverage D_{mono} for edge-aware filtering, we compute the discrepancies between the gradients of the two maps: $\Delta(p_{\text{grad}}) = |\frac{|\nabla D_{\text{mono}}|}{D_{\text{mono}}} - \frac{|\nabla D'_{\text{geom}}|}{D'_{\text{geom}}}|$ and discard pixels whose discrepancies exceed a predefined threshold τ_{grad}. This approach effectively filters both bleeding artifacts and noisy transient objects without relying on manual category lists, as depicted in Fig. 3.

3.3. Dataset Statistics

In summary, we identify 2,474 candidate scenes from MegaScenes with more than 100 registered images. Of these, 609 scenes are filtered out due to dynamic content, reconstruction errors, or geometric inconsistencies. Our final MD-X dataset comprises 1,865 reconstructions totaling 466k images. We reserve 127 scenes for testing, providing a novel set for evaluating both pretrained and fine-tuned methods. A comparison table with MegaDepth is provided in the supplementary.

4. Simulating Long-Tail Scenes

With MD-X providing reliable 3D supervision, the remaining challenge is a complementary supervision coverage problem: existing 3D foundation models are trained predominantly on the head of the Internet-photo distribution, where image collections are large, redundant, and visually well-connected. In this regime, models can rely on strong covisibility and abundant local correspondences. However, most real Internet photo collections lie in the long tail, where views are sparse, unevenly distributed, and only weakly connected. A more complete 3D prior should therefore be robust not only to diverse scene content, but also to this underrepresented observation regime. Rather than seeking unreliable supervision from true long-tail scenes, we start from well-reconstructed scenes in MD-X and sample subsets whose covisibility structure matches that of real long-tail collections. In this way, we expose the model to the missing part of the training distribution while inheriting trustworthy 3D supervision from the full reconstruction.

4.1. Defining Properties of Long-Tail Scenes

Common issues like transient occluders and motion blur affect Internet photos broadly, but they are not the primary bottleneck for long-tail scenes. The more fundamental challenge lies in their viewpoint distribution. In these scenes, sparse camera placements lead to limited mutual overlap between images. This results in fragmented, weakly connected clusters rather than a cohesive set, which poses a major hurdle for reliable 3D reconstruction. Because accurate camera poses are often unavailable for such scenes, we characterize this regime using statistics of the SfM view graph rather than absolute camera geometry. Our analysis reveals two consistent patterns: (1) *sparser connectivity*: scenes with low registration rates (e.g., only 20% of images registered) contain a substantially larger fraction of low-degree nodes, with 8% of cameras having degree two or less, compared with only 3% in well-reconstructed head scenes. This indicates that cameras in long-tail scenes are poorly connected, forming fragmented clusters with limited covisibility. (2) *weaker connections*: even among connected image pairs, the average number of geometrically verified feature matches is significantly lower in long-tail scenes than in head scenes (294.8 vs. 395.3), indicating reduced overlap and weaker geometric consistency.[1] Together, these observations show that the long tail is not simply a regime of fewer images, but one of sparse and weakly connected observation graphs.

Based on these findings, our sampling process should satisfy three requirements:

- **Viewpoint Diversity:** The sampled views should cover a wide range of viewing directions, ensuring that emulated scenes span diverse visual perspectives.
- **Sparsity:** The selected views should be far enough apart to mimic the wide baselines typical of long-tail scenes, *e.g.* loosely connected views or views from disconnected scene components, encouraging the model to learn robust geo-

[1]To avoid statistics being dominated by severely noisy scenes, we compute these measurements only on long-tail subsets containing at least five registered images.

metric priors rather than relying on dense feature matches.

- **Local Reconstructability:** Despite the sparsity, views within each sampled scene component should retain enough covisibility to remain locally reconstructable, since zero-overlap samples within a scene component can lead to unstable training signals and difficult optimization.

4.2. Sparsity-Aware Sampling Strategy

We therefore formulate the sampling task as sampling N views that form at most N_{cc} connected components, in order to emulate a long-tail scene with multiple weakly connected or disconnected scene components. Specifically, components are allowed to be disconnected from one another, but within each sampled component we still require sufficient internal covisibility for local reconstructability. We find that naïve random or uniform subsampling often fails to satisfy this balance, producing either zero-overlap sets within scene components or clusters biased toward dense regions. We instead propose a structured sampling process. We first partition views into strongly connected communities and then select a minimal yet diverse subset that ensures both community coverage and global connectivity. This process is illustrated in Fig. 4.

Graph Communities. To promote viewpoint diversity in our sampling, we first identify the dominant "viewing areas" within each scene. We represent the SfM structure as a view graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a camera view and each edge $(v_i, v_j) \in E$ is weighted by the number of feature matches w_{ij}. We prune edges with $w_{ij} < 50$ to remove minor overlaps, resulting in a filtered graph $G' = (V, E')$ that preserves only meaningful covisibility relationships. To reveal clusters of cameras with dense internal connectivity, we perform community detection (e.g., Louvain community detection [4]) on the view graph. This yields viewpoint groups C_k that efficiently capture distinct visual regions and the dominant perspectives of the scene. We then randomly partition the graph into N_{cc} connected components that span different communities and do the following steps *within each graph partition*. The partition algorithm is provided in the supplementary material.

Minimal Connectivity Subgraph. To preserve overall scene connectivity while maintaining sparsity and view diversity within limited nodes, we construct a minimal structure linking all identified communities without reintroducing dense redundancy within each partition. We then compute an approximate Steiner tree to link all of these nodes [12, 16].[2] In particular, for each training batch for a given training scene, we first randomly select one representative view $v_k \in C_k$ from each community C_k to form the terminal set $T = \{v_k\}$. An approximate Steiner tree

[2]A Steiner tree aims to span a specified set of *terminal* nodes while introducing only the minimal set of intermediate nodes required for connectivity.

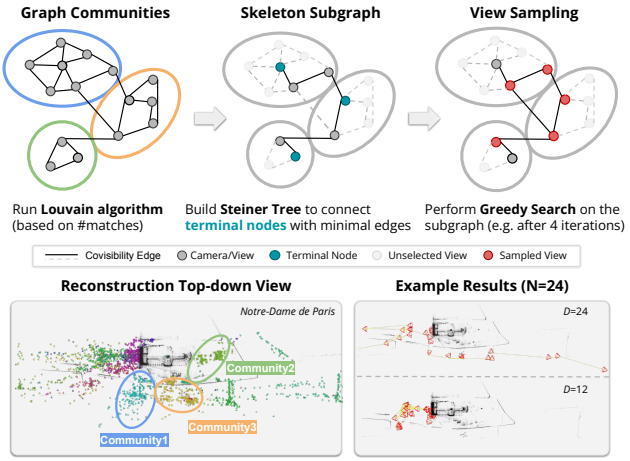

Figure 4. **Sparsity-aware sampling strategy. Top:** Our method follows a multi-stage process: (1) Apply the *Louvain algorithm* to the view graph to identify distinct viewpoint communities. (2) From each community, randomly select a terminal view and construct an approximate *Steiner Tree* to form a minimal, connected subgraph spanning these communities. (3) Perform a *Greedy Search* on this subgraph to select a sparse and diverse set of views. This procedure aims to cover as many communities as possible while ensuring a wide spatial distribution of cameras within each community. **Bottom:** A *search depth* parameter controls the final view coverage. In this example, we sample $N = 24$ views from the scene with $N_{cc} = 1$. With search depth $D = 24$, all views are selected via greedy search, producing a more evenly spread distribution. With $D = 12$, 12 views come from greedy search and the remaining 12 are sampled locally from the neighborhoods of selected nodes, resulting in a more concentrated distribution.

algorithm then constructs a minimal connected subgraph $G_{\text{sub}} = (V_{\text{sub}}, E_{\text{sub}}), \quad T \subseteq V_{\text{sub}} \subseteq V$, that spans all terminal nodes using only the necessary intermediate nodes. This yields a compact subgraph connecting all communities using the fewest necessary nodes and edges, preserving global consistency while retaining sparsity. Since G_{sub} can have an arbitrary number of nodes, we need to perform additional sampling to get desired number of views for the training and testing batches.

Greedy View Sampling. Inspired by skeletal sets [26], we perform greedy view sampling on the subgraph G_{sub} to select a diverse subset of views for long-tail emulation. The objective is to iteratively expand the sampled set toward broad spatial coverage while maintaining sufficient covisibility among selected view pairs.

At each iteration, the algorithm aims to select the next view based on two criteria: (1) *Community novelty*: preferring cameras that belong to previously unseen communities, thereby introducing new viewing directions and reducing redundancy; and (2) *Spatial distance*: encouraging selection of cameras farther from the current viewpoint to promote wider baseline coverage. Specifically, the algorithm operates on a current node v and its connected neighborhood N_v. Let S denote the set of already sampled nodes and M be the community map. We first determine which communities have already been reached in S, form-

ing the set $S_{\text{comm}} = \{M[s] \mid s \in S\}$. For each neighbor $u \in N_v$, we then evaluate its community novelty by checking whether $M[u] \notin S_{\text{comm}}$, and compute its spatial distance as $\|\text{Pos}(u) - \text{Pos}(v)\|_2$, where $\text{Pos}(\cdot)$ is camera position. Details for this algorithm are provided in the supplemental material. All candidate neighbors are ranked lexicographically by these two attributes, and the top-ranked neighbor u^* is selected as the next sampled node. This procedure is repeated for D iterations (i.e., the search depth).

Implementation. In practice, we compute a fixed set of communities $\mathcal{C} = \{C_k\}$ for each scene. To form a training batch of N images for a scene, we first randomly divide the N samples across all N_{cc} partitions. In each partition, greedy view sampling stops once either a predefined search-depth limit D is reached or the target number of views assigned to that partition has been sampled. Here, D controls how far the search expands within a partition, hence the sparsity of the resulting set. If this process still produces fewer than N nodes in total, we fill the remaining slots by randomly sampling nodes from the local neighborhoods of the previously sampled nodes. Fig. 4 illustrates an example in which $N = 24$ and $N_{cc} = 1$, and shows the different sparsities of the sampled set obtained under different values of D. Before training, we run the proposed sampling algorithm offline to generate mini-batches of 24 nodes, avoiding costly graph loading during training. We then perform depth-first search from random seed nodes to subsample 2 to 24 images for training batches.

5. Experiments

We evaluate how our approach improves 3D reconstruction in the long-tail regime of Internet photo collections. First, we show quantitative results on the proposed MD-X benchmark, demonstrating qualitative improvements on real-world long-tail and doppelganger scenes. We then analyze the effect of the proposed dataset and sampling strategy, and finally verify that our fine-tuned models preserve strong performance on standard, curated benchmarks. Further implementation details and additional results are in the supplementary material.

5.1. Experimental Setup

Backbones and variants. We finetune two feed-forward 3D foundation models, π^3 [34] and VGGT [30], on MD-X using our proposed sampling strategy. We adopt the loss functions from π^3 [34] and VGGT [30]. To preserve pretrained geometric fidelity, we finetune only the Alternating-Attention modules and keep the point cloud and camera decoders frozen. More training details are in the supplementary. The resulting models are denoted as π^3-FT and VGGT-FT.

To study how our proposed view sampling strategy affects performance, we finetune π^3 on clean Internet data using four sampling schemes:

- DENSE: training batches with densely overlapping views where $D = 5$ and $N_{cc} = 1$,
- SPARSE: long-tail–like sampling emphasizing wide baselines where $D = 24$ and $N_{cc} = 4$,
- MIXED: a combination of dense and sparse batches for balanced learning with $D \in [5, 24]$ and $N_{cc} \in [1, 4]$,
- RANDOM: random view sampling.

Unless otherwise noted, FT (e.g., π^3-FT) refers to the model finetuned on the cleaned dataset using the MIXED sampling strategy above. We additionally train a DIRTY variant on Internet data (using the same Mixed scheme) without the filtering strategy in Sec. 3.1, while keeping the same depth refinement pipeline in Sec. 3.2, to assess robustness to label noise and data contamination.

Table 1. **Quantitative results on MegaDepth-X** for camera pose and point map estimation across two difficulty levels. Our finetuned models (π^3-FT and VGGT-FT) trained with the proposed dataset and sampling strategy consistently outperform pretrained baselines, especially on harder, sparser scenes.

	Method	Camera Pose Estimation					Point Map Estimation					
							Acc↓		Comp↓		NC↑	
		RRA@5↑	RTA@5↑	AUC@5↑	MRE↓	MTE↓	Mean	Med.	Mean	Med.	Mean	Med.
easy	π^3	88.97	68.79	45.84	4.12	7.82	0.055	0.030	0.039	0.019	0.712	0.822
	π^3-FT	**95.64**	**76.85**	**55.58**	**1.64**	**5.50**	**0.035**	**0.020**	**0.024**	**0.012**	**0.724**	**0.837**
	VGGT	84.17	58.47	35.32	4.55	9.93	0.093	0.047	0.055	0.026	0.695	0.798
	VGGT-FT	**92.41**	**71.12**	**48.78**	**2.70**	**7.02**	**0.050**	**0.027**	**0.033**	**0.014**	**0.719**	**0.833**
hard	π^3	75.31	59.16	36.93	12.21	10.82	0.101	0.065	0.133	0.090	0.689	0.786
	π^3-FT	**86.40**	**71.00**	**47.93**	**5.72**	**7.27**	**0.068**	**0.041**	**0.066**	**0.041**	**0.713**	**0.818**
	VGGT	70.98	52.98	29.10	13.20	13.34	0.149	0.092	0.151	0.104	0.675	0.764
	VGGT-FT	**81.07**	**65.59**	**41.49**	**7.22**	**9.05**	**0.089**	**0.053**	**0.084**	**0.055**	**0.709**	**0.814**

Evaluation Metrics. For camera pose estimation, we follow prior work [30, 34] and report Relative Rotation Accuracy (RRA), Relative Translation Accuracy (RTA), and their combined Area Under Curve (AUC). We also report mean rotation and translation errors (MRE and MTE, in degrees). For point map evaluation, we follow prior work [2, 29, 31, 33, 34] and report Accuracy (Acc), Completeness (Comp), and Normal Consistency (NC), each computed as the mean and median across test scenes.

5.2. Internet Photo Evaluation

We first evaluate models on the proposed MD-X benchmark, which contains Internet photo collections of varying sparsity and difficulty. For each test scene, we sample 24 images from the reconstructed scene graph using our sampling algorithm, and categorize them into *easy* ($D = 5$, $N_{cc} = 1$) and *hard* ($D = 24$, $N_{cc} = 4$) subsets according to the greedy search depth used for test data sampling.

Quantitative Results. Tab. 1 reports quantitative results for camera pose and point map estimation across three difficulty levels on MD-X. Finetuning markedly improves both π^3 and VGGT over their pretrained baselines, with larger gains observed in harder, sparser scenes. These improvements hold across metrics indicate that the fine-tuned models better capture global structure and maintain consistent 3D geometry

Table 2. **Ablation study on MegaDepth-X.** Finetuning on the cleaned dataset with MIXED dense–sparse sampling (π^3-FT) yields the best overall performance, while training on unfiltered data (DIRTY) degrades accuracy.

	Method	Camera Pose Estimation					Point Map Estimation					
							Acc↓		Comp↓		NC↑	
		RRA@5↑	RTA@5↑	AUC@5↑	MRE↓	MTE↓	Mean	Med.	Mean	Med.	Mean	Med.
easy	π^3	88.97	68.79	45.84	4.12	7.82	0.055	0.030	0.039	0.019	0.712	0.822
	π^3-FT	95.64	**76.85**	55.58	1.64	**5.50**	**0.035**	**0.020**	**0.024**	**0.012**	0.724	**0.837**
	π^3-DIRTY	91.25	72.80	51.77	5.16	7.28	0.075	0.052	0.081	0.051	0.710	0.818
	π^3-RANDOM	95.08	76.42	55.00	1.78	5.72	0.039	0.021	0.026	0.013	0.720	0.831
	π^3-DENSE	95.13	76.73	**55.65**	1.84	5.61	0.036	**0.020**	0.026	0.013	**0.725**	**0.837**
	π^3-SPARSE	**96.27**	76.46	55.12	**1.61**	5.59	0.038	**0.020**	0.026	0.013	0.723	0.835
hard	π^3	75.31	59.16	36.93	12.21	10.82	0.101	0.065	0.133	0.090	0.689	0.786
	π^3-FT	**86.40**	**71.00**	**47.93**	**5.72**	**7.27**	**0.068**	0.041	0.066	0.041	**0.713**	**0.818**
	π^3-DIRTY	81.10	65.99	43.74	11.86	9.72	0.130	0.094	0.139	0.091	0.693	0.791
	π^3-RANDOM	85.93	69.84	47.17	6.53	7.78	0.071	**0.040**	0.073	0.045	0.708	0.812
	π^3-DENSE	85.82	70.06	47.47	6.04	7.64	0.071	0.042	**0.062**	**0.035**	**0.713**	0.817
	π^3-SPARSE	85.97	70.53	47.13	6.05	7.52	0.070	**0.040**	0.070	0.041	0.710	0.814

in sparse settings.

Ablation Analysis. We analyze the effects of data quality and sampling strategies, with results shown in Tab. 2. Training on unfiltered (DIRTY) data consistently reduces accuracy, even performing worse than the pretrained model in point-map estimation on both the *easy* and *hard* levels, highlighting the importance of clean supervision for robust generalization. Among sampling schemes, RANDOM sampling yields reasonable camera pose accuracy but provides limited improvement in point map reconstruction, emphasizing the importance of adequate covisibilities in training batches. DENSE sampling performs well on easier scenes but is less effective under sparse conditions. SPARSE sampling alone does not yield the best trade-off. Although it exposes the model to more challenging cases, MIXED sampling achieves slightly better overall performance across difficulty levels.

Qualitative Analysis. We show qualitative results for three settings: the MD-X test set, real-world long-tail Internet scenes, and doppelganger scenes.

MegaDepth-X Visualization. Fig. 5 shows reconstruction results on the MD-X test set across *easy* and *hard* levels. Our fine-tuned model produces more accurate camera poses, more dense and consistent 3D point maps compared to the pretrained baseline, especially on sparse (*hard*) scenes. It generalizes well across varying camera intrinsics and challenging appearance changes such as day-night shifts.

Real Long-Tail Scenes. Real long-tail Internet scenes often contain fewer than 100 usable photos captured from uneven viewpoints and mixed with transient or irrelevant content. Classical SfM pipelines, e.g., COLMAP, typically fail to register most images, producing extremely sparse geometry or incomplete reconstructions. Pretrained models struggle under these conditions, yielding low-confidence predictions and fragmented structures. Our finetuned model remains stable and reconstructs coherent global geometry. As shown in Fig. 6, our model successfully reconstructs dense geometry from very few views, and handles doppelganger ambiguities with higher confidence, demonstrating strong robustness and generalization to real-world long-tail scenes. In the supplementary material, we provide more results on doppelganger scenes.

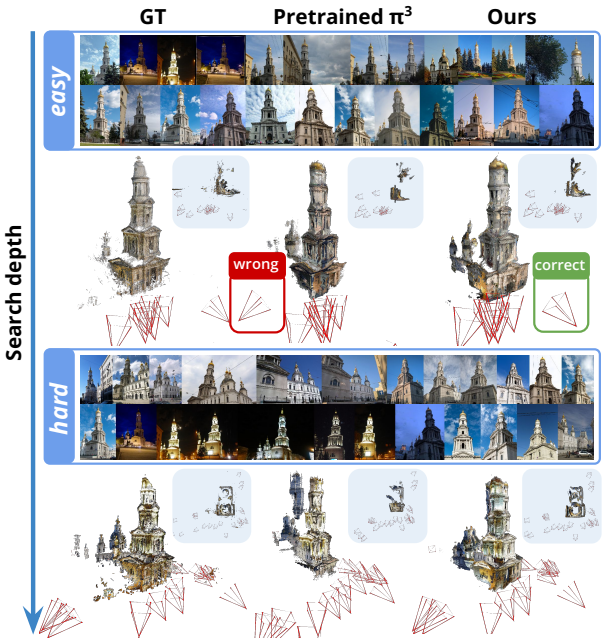

Figure 5. **Reconstruction results on the MegaDepth-X test set across two difficulty levels.** For each level, the top row shows the full 24-image input set, and the bottom row compares reconstructions from ground truth, pretrained π^3, and our finetuned model with top-down views shown in the insets. Our model shows clearer improvements in the *hard* setting, where the inputs are more challenging. Note that *hard* was obtained using a deeper search depth than *easy*.

Table 3. **Camera pose estimation on RealEstate10K [38] and CO3Dv2 [17].** We follow π^3's pose sampling conventions. Our fine-tuned models, trained on proposed Internet data dataset, remain comparable to pretrained baselines, demonstrating generalization to standard benchmarks.

Method	RealEstate10K					CO3Dv2				
	RRA@5↑	RTA@5↑	AUC@5↑	MRE↓	MTE↓	RRA@5↑	RTA@5↑	AUC@5↑	MRE↓	MTE↓
π^3	98.79	**79.61**	**62.82**	**0.51**	**5.65**	93.24	84.47	57.12	3.04	4.28
π^3-FT	**98.80**	77.78	60.01	**0.51**	6.13	**93.97**	**84.50**	**57.61**	**2.96**	**4.26**
VGGT	97.49	62.32	38.09	1.03	8.66	96.97	86.19	**67.84**	2.33	3.95
VGGT-FT	**98.23**	**71.88**	**48.23**	**0.82**	**6.85**	**97.11**	**86.27**	67.81	**2.29**	**3.92**

5.3. Generalization to Standard Benchmarks

We next examine whether the finetuned models preserve generalization on standard, curated benchmarks.

Relative Pose Estimation. We evaluate on RealEstate-10K [38] and CO3Dv2 [17], following π^3's pose sampling conventions. As shown in Tab. 3, fine-tuning on Internet data generally maintains the performance of both backbones, and yields modest improvements for VGGT in particular. These results indicate that robustness learned from sparse, in-the-wild Internet photos does not compromise generalization to standard 3D benchmarks.

Point Map Estimation. Results on DTU [9], ETH3D [23], 7-Scenes [24], and NRGBD [2] (Tab. 4&5) show that our model maintains comparable reconstruction accuracy on

Figure 6. **Reconstruction results on real long-tail Internet scenes.** Each scene contains only a handful of photos with uneven viewpoints and noisy content, where COLMAP fails to register most images and produces extremely sparse geometry. Pretrained π^3 makes low-confidence predictions and incomplete reconstructions, while our fine-tuned model discovers the correct large-scale layout (e.g., (1) *Novo-Znamenka Manor*, 66 images, 13 registered), handles very few-view inputs and recovers dense geometry ((2) *Sobanski Palace in Guzow*, 95 images, 11 registered), reconstructs more complete structures under sparse, long-tail settings ((3) *Delizia del Verginese (Gambulaga, Portomaggiore)*, 69 images, 11 registered, (5) *Chitharal Jain Monuments*, 44 images, 15 registered), resolves doppelganger ambiguity ((4) *Hoshang's Tomb*, 85 images, 40 registered), and even works when COLMAP completely fails ((6) *Chapel of Saint Andrew's cathedral (Saint Petersburg)*, 94 images, 0 registered). These results demonstrate that our model remains robust and confident under severe sparsity and ambiguity in real long-tail Internet scenes. **For each scene, the confidence threshold is the same for pretrained π^3 and our method.**

Table 4. **Point map estimation on DTU [9] and ETH3D [23].** Finetuning on the proposed Internet photo dataset retain overall reconstruction quality on DTU, while performance on ETH3D decreases due to domain mismatch with Internet imagery. These results show that the model adapts to Internet photos without drifting too much on out-of-domain benchmarks.

Method	DTU						ETH3D					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
π^3	**1.151**	**0.622**	**1.793**	0.629	**0.668**	**0.754**	**0.188**	**0.126**	**0.211**	**0.129**	**0.872**	**0.967**
π^3-FT	1.202	0.642	1.928	**0.593**	0.666	0.751	0.199	0.142	0.242	0.151	0.861	0.955
VGGT	1.308	0.761	1.929	1.015	0.665	0.750	**0.270**	**0.174**	**0.304**	**0.180**	**0.841**	**0.942**
VGGT-FT	**1.283**	**0.759**	**1.900**	**0.953**	**0.669**	**0.756**	0.282	0.205	0.394	0.225	0.838	0.927

Table 5. **Point map estimation on 7-Scenes [24] and NRGBD [2] datasets.** We evaluate both sparse-view and dense-view settings. Finetuning on Internet photos yields comparable performance to pretrained baselines with minor variations, indicating our method preserves generalization across diverse real world and synthetic datasets.

View	Method	7-Scenes						NRGBD					
		Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
sparse	π^3	0.047	0.029	0.074	0.049	**0.741**	0.840	**0.024**	**0.013**	**0.028**	**0.013**	**0.909**	**0.991**
	π^3-FT	**0.046**	**0.027**	**0.072**	**0.046**	0.739	**0.841**	**0.024**	0.014	**0.028**	0.014	0.903	0.990
	VGGT	**0.044**	**0.024**	**0.056**	**0.033**	0.733	**0.846**	**0.049**	**0.027**	**0.066**	**0.037**	**0.882**	**0.979**
	VGGT-FT	0.062	0.046	0.097	0.070	**0.738**	0.844	0.071	0.046	0.071	0.041	0.875	0.959
dense	π^3	**0.016**	**0.007**	**0.022**	**0.011**	**0.689**	**0.792**	**0.013**	**0.007**	**0.014**	0.006	**0.874**	**0.981**
	π^3-FT	**0.016**	**0.007**	0.023	**0.011**	0.686	0.789	**0.013**	**0.007**	**0.014**	**0.005**	0.864	0.978
	VGGT	0.022	0.008	**0.026**	**0.012**	0.667	0.760	**0.015**	**0.008**	**0.015**	**0.006**	**0.871**	**0.982**
	VGGT-FT	**0.016**	**0.007**	0.027	**0.012**	**0.681**	**0.781**	**0.015**	0.008	0.016	**0.006**	0.859	0.981

DTU, 7-Scenes and NRGBD. We observe a performance decrease on ETH3D and a mild drop for VGGT under sparse NRGBD, likely reflecting the domain gap between these clean, controlled datasets and Internet imagery. Overall, the results indicate that training on diverse Internet photos preserves cross-dataset generalization without overfitting.

6. Conclusion

We presented a step towards robust, Internet-scale 3D reconstruction by defining and addressing the long-tail regime of Internet photo collections. Through the MegaDepth-X dataset and a sparsity-aware sampling strategy, we augment the ability of 3D foundation models to recover consistent geometry from sparse, noisy, and ambiguous imagery, where classical SfM and SOTA feed-forward 3D reconstruction models fail, and demonstrates disambiguation of doppelganger scenes while maintaining generalization across benchmarks.

Our dataset currently focuses on landmark-scale scenes, representing only a small fraction of the landscape of Internet photos. Bootstrapping on the current dataset and refining models for reconstructions of even more longed-tail data remains an important direction for future work. Extending this framework beyond landmarks to everyday objects, indoor scenes, and other Internet photo domains offers a promising path toward a truly universal 3D foundation model.

Acknowledgments This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project). We thank Joseph Tung, Yiwen Zhang, Hanyu Chen and Haian Jin for discussion and help with MegaScenes dataset and depth post-processing.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54(10): 105–112, 2011. 2

[2] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 6, 7, 8

[3] Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbach-Elor. Extreme rotation estimation in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1061–1070, 2025. 3

[4] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008. 5

[5] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 3

[6] 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/CVF International Conference on Computer Vision*, pages 34–44, 2023. 3

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 3

[8] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, and Marc Pollefeys. Building Rome on a Cloudless Day. In *ECCV*, 2010. 2

[9] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 7, 8

[10] Hanwen Jiang, Hanwen Jiang, Arjun Karpur, Bingyi Cao, Qixing Huang, and Qi-Xing Huang. Omniglue: Generalizable feature matching with foundation model guidance. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19865–19875, 2024. 3

[11] Arjun Karpur, Guilherme Perrotta, Ricardo Martin-Brualla, Howard Zhou, and Andre F. de Araújo. Lfm-3d: Learnable feature matching across wide baselines using 3d signals. *2024 International Conference on 3D Vision (3DV)*, pages 11–20, 2023. 3

[12] Lawrence Kou, George Markowsky, and Leonard Berman. A fast algorithm for steiner trees. *Acta informatica*, 15(2): 141–145, 1981. 5

[13] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 2, 3

[14] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2041–2050, 2018. 2, 3

[15] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching atokens are frozen. *arXiv preprint arXiv:2306.13643*, 2023. 3

[16] Kurt Mehlhorn. A faster approximation algorithm for the steiner problem in graphs. *Information Processing Letters*, 27 (3):125–128, 1988. 5

[17] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10881–10891, 2021. 7

[18] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3

[19] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3

[20] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 3

[21] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 2

[22] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[23] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, 2017. 7, 8

[24] 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*, pages 2930–2937, 2013. 7, 8

[25] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[26] Noah Snavely, Steven M Seitz, and Richard Szeliski. Skeletal graphs for efficient structure from motion. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 5

[27] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 1, 2, 3

[28] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020. 3

[29] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 6

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 6

[31] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 6

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

[33] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023. 2, 6

[34] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 1, 2, 6

[35] Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features, 2025. 3

[36] 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 Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2

[37] 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 Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025. 2

[38] 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. 7