Trainspotting: Learning Object-Centric 3D Reconstruction from Found Photos

Hanyu Chen Noah Snavely
Cornell University

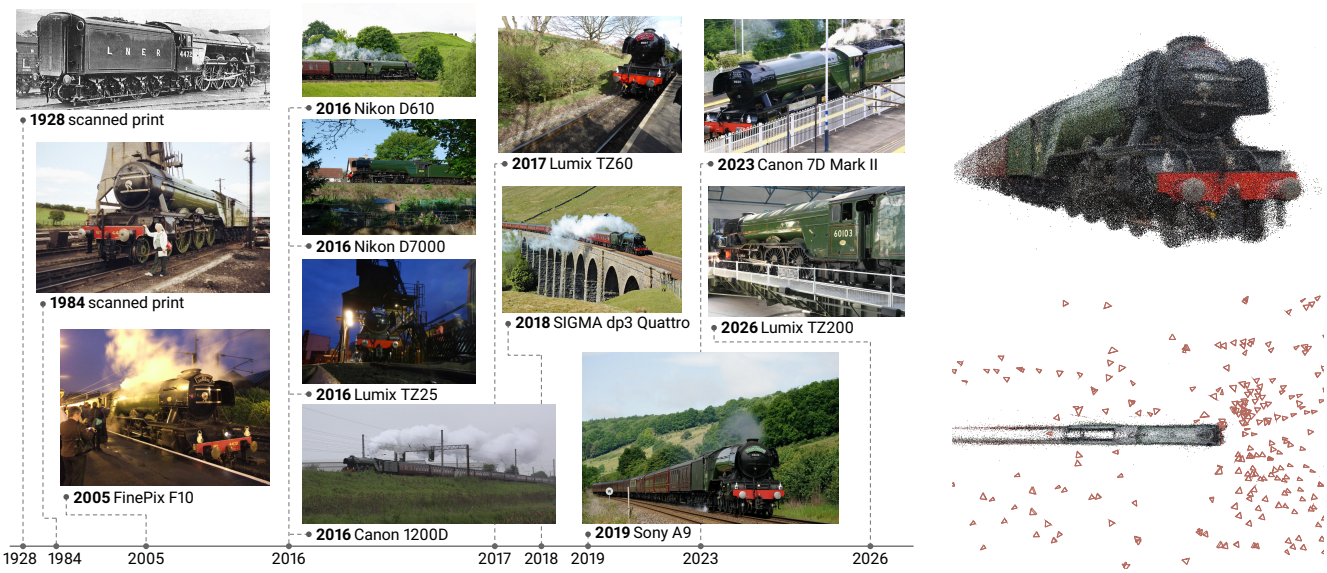


Figure 1. **One object over one century.** Wikimedia Commons holds 727 photos of the locomotive *Flying Scotsman*, taken from the 1920s to 2026 by 187 photographers using 123 camera models. The twelve photos on the left are labeled with their year and camera; our recovered sparse point cloud and camera poses are shown on the right. TRAINSPOTTING contains 12,068 such objects reconstructed from 311,886 photos from Wikimedia Commons. We use this dataset to substantially improve feedforward 3D reconstruction models on in-the-wild object-centric photo collections.

Abstract

State-of-the-art feedforward reconstruction models perform poorly on in-the-wild object-centric photo collections. They are trained largely on single-session object captures whose images vary in viewpoint while the object and its environment remain nearly fixed. Public photo archives contain a different form of multi-view data. They hold found photos *of the same object or design: photos taken independently by many people across years, environments, cameras, and sometimes physical units. In this work, we introduce* TRAINSPOTTING*, a dataset of 12,068 objects and 311,886 photos from Wikimedia Commons. We reconstruct each object to recover camera poses, multi-view stereo depth, and object masks. We categorize the objects into three regimes: objects fixed at one site, objects that move between sites, and designs shared by many physical units. Evaluating recent feedforward reconstruction models on* TRAINSPOTTING *shows that they fail most severely in the last two regimes. Fine-tuning VGGT-Ω on* TRAINSPOTTING *significantly improves camera pose, depth, and 3D geometry on held-out objects. We observe similar improvements on the independent NAVI in-the-wild benchmark, which contains photos of each object in different environments. Performance on conventional benchmarks remains largely unchanged.*

1. Introduction

Choose life. Choose a train, a statue, a pipe organ. Choose a 1967 Volkswagen Beetle, a K6 red telephone box in London. We chose not to choose. We reconstructed them all.

Found photos of objects like those listed above accumulate in public archives such as Wikimedia Commons over many years. Feedforward reconstruction models [48, 49, 52], however, obtain most of their object-level training data from datasets that record each object in a single video or capture session [28, 31, 55]. These captures densely sample nearby views while the object and its environment remain nearly fixed.

The photos of an object on Wikimedia Commons follow no shared capture plan. Their viewpoints are unevenly sampled,



Figure 2. **VGGT-Ω fails on the *Buick Riviera*.** We visualize the camera poses and point map predicted by VGGT-Ω from sixteen photos of the Buick Riviera (multiple physical instances of the same car design). The views are misregistered and the car is reconstructed as scattered fragments. Figure 6 shows the reconstruction of our fine-tuned model from the same photos.

and their environment may change dramatically from one photo to the next (e.g., a ship on the open seas vs. the same ship in port). An individual object may be repainted or rebuilt over time, and photos of a mass-produced product, like a specific make and model of car, may depict different physical instances. Datasets of *landmarks* and other places have been built from similar in-the-wild Internet photo collections [24, 38, 42]. However, these scene-level photo collections depict a largely consistent physical space, whereas photos of an object often share only the object itself in many different contexts and times. Figure 1 shows an example of such photos, depicting the locomotive *Flying Scotsman* captured over a century in many different locations.

Standard reconstruction pipelines are unreliable in such in-the-wild settings, and so are feedforward reconstruction models, which are largely trained on their output (Figure 2). Classical structure from motion relies on repeatable correspondences that agree with one rigid scene [35]. A changing environment provides few such correspondences, and the object may occupy only a small part of each photo. Changes to an individual object over time and differences between physical units further weaken geometric consistency. Together, these factors can result in incorrect reconstructions, so building a dataset from these photos requires identifying the photos of the intended object, matching them robustly, and verifying the recovered geometry.

In this work, we propose a dataset and an evaluation protocol for feedforward reconstruction of objects from in-the-wild photo collections. To build this dataset, we collect candidate categories from Wikidata and Wikimedia Commons, such as *Cristo Redentor*, the *Hubble Space Telescope*, and the *Ford Mustang*. We use a large language model (LLM) to retain the categories that depict a rigid physical object or design (Section 3). To address the failures of standard pipelines on these collections, we propose a new object reconstruction pipeline, combining dense learned matching with geometric verification of the recovered camera poses and depth maps (Section 4). The result of our pipeline is TRAINSPOTTING, an in-the-wild object-centric dataset that contains 12,068 objects and 311,886 photos from Wikimedia Commons paired with camera poses, MVS depth, and object masks. An example reconstruction of the *Flying Scotsman* is shown on the right of Figure 1.

Table 1. **Real-image datasets for 3D reconstruction tasks.** Datasets above the line record each object or scene in a single capture within minutes, while datasets below contain photos of each object or scene taken by many people spanning days to years. NAVI and Pix3D align images to 3D models by hand; all other datasets recover poses with structure from motion.

Dataset	Type	Count	Images	Span	Depth
OmniObject3D [53]	objects	6k	1.2M	minutes	–
CO3D [31]	objects	19k	1.5M	minutes	MVS
MVImgNet2.0 [54]	objects	520k	15.5M	minutes	–
uCO3D [28]	objects	170k	34M	minutes	mono.
NAVI [19]	objects	36	10.5k	days	scan
Pix3D [40]	objects	395	10k	days/years	scan/CAD
MegaDepth [24]	scenes	196	130k	years	MVS
MegaDepth-X [23]	scenes	1.9k	440k	years	MVS
MegaScenes [42]	scenes	80k	2M	years	–
TRAINSPOTTING (ours)	objects	12k	312k	years	MVS

We evaluate recent feedforward reconstruction models on TRAINSPOTTING and find that they perform substantially worse on these in-the-wild collections than on conventional benchmarks. Therefore, we fine-tune one such model, VGGT-Ω, on TRAINSPOTTING and show that the resulting model, SPOTTER, significantly improves camera pose, depth, and 3D geometry on these objects, raising camera-pose AUC@30 from 25 to 81 on held-out objects and from 19 to 76 on NAVI in-the-wild [19], an independent benchmark of objects photographed across different environments, while performance on conventional benchmarks is largely retained. Together, these results suggest that TRAINSPOTTING can help feedforward reconstruction models generalize to the real-world diversity of object photos.

2. Related Work

3D datasets. 3D datasets have grown rapidly in scale. Synthetic datasets such as Objaverse [9, 10] and ShapeNet [7] provide textured 3D models. Datasets built with 3D scanners include DTU [1], Google Scanned Objects [12], and OmniObject3D [53], and contain at most a few thousand objects. Meanwhile, CO3D [31], MVImgNet [54, 62], and uCO3D [28] reach a larger scale by crowdsourcing videos of objects and annotating them with 3D information via structure from motion. In such datasets, each object is recorded in one coordinated capture [28, 31, 55, 62]. ScanNet [8], ScanNet++ [61], and DL3DV [27] record scenes in the same way. Few datasets photograph objects in different environments. PASCAL3D+ [56], ObjectNet3D [57], and Pix3D [40] align CAD models to internet images of objects, but the models only approximate the depicted objects. NAVI [19] is the closest dataset to our setting. It photographs each object in different environments with different cameras and aligns each image to a 3D scan of that object. In all of these object-scale datasets, each image is aligned to its corresponding 3D model by hand, and per-image geometry such as depth is rendered from that model rather than reconstructed from the photos. In contrast, TRAINSPOTTING consists of multi-view photo collections of objects taken by

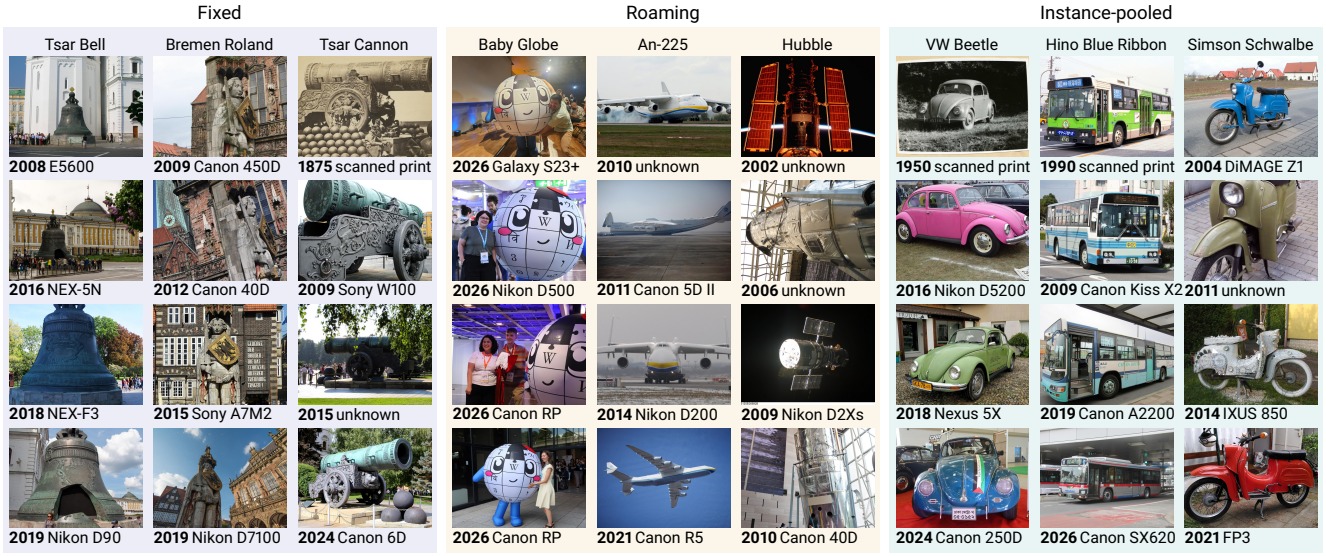


Figure 3. **Examples of capture variation.** Each column displays photos of example objects, grouped into the three regimes (fixed, roaming, and instance-pooled) and labeled with their year and camera. *Fixed* objects stay at one site, *roaming* objects are single instances that move between sites, and *instance-pooled* objects are different physical instances of one shared design.

different people over many years, with camera poses and depth recovered by an automated reconstruction pipeline, and is much larger than NAVI. Table 1 compares TRAINSPOTTING with other datasets of real images used for 3D reconstruction.

Reconstruction from internet photo collections. Photo Tourism [38] and Building Rome in a Day [2] reconstructed landmarks from unordered photos taken by many photographers. Later datasets use such reconstructions as supervision for learned models. MegaDepth [24] reconstructs about 200 landmarks from photos on Flickr with COLMAP [35], filters the resulting multi-view stereo depth, and uses it to supervise single-view depth prediction. MegaScenes [42] reconstructs 80k Wikimedia Commons categories of scenes with COLMAP using SIFT features. MegaDepth-X [23] selects densely photographed MegaScenes scenes and reconstructs them with MASt3R-SfM [13] and doppelganger disambiguation [58]. All three contain scene-level rather than object-centric reconstructions. Of the 12,068 objects in TRAINSPOTTING, only 1,163 (9.6%) are reconstructed as scenes in MegaScenes and 56 (0.5%) in MegaDepth-X, almost all of them objects permanently installed at one site, such as statues and memorials. Objects that move between sites or exist as many physical units are almost entirely absent from these datasets.

Learned matching and structure from motion. Classical structure from motion (SfM) methods like COLMAP [35] match hand-crafted keypoints and register images incrementally. SuperPoint [11] and SuperGlue [34] replace traditional keypoint detection and matching with learned components. LoFTR [39] proposes detector-free matching, predicting matches directly with a transformer. Dense matchers such as DKM [14], RoMa [15], and RoMa v2 [16] predict a warp for every pixel of an image pair, aiming to recover correspondences under large changes in viewpoint and illumination. Several SfM systems, such as Detector-Free SfM [18] and Dense-SfM [21], build on such dense matches to form multi-view tracks and refine the tracks together with the geometry. MASt3R-SfM [13] builds on the matches of MASt3R [22] and jointly aligns all images. VGGSfM [47] learns the entire reconstruction pipeline, predicting tracks, cameras, and points and refining them with differentiable bundle adjustment. We keep the incremental reconstruction of COLMAP and supply it with multi-view tracks built from RoMa v2 matches that pass confidence and geometric checks (Section 4).

Feedforward 3D reconstruction. Feedforward models that regress cameras, depth, and point maps directly from a set of images have become a standard approach to 3D reconstruction [20, 22, 25, 41, 48, 49, 51, 52, 59, 64]. Most of their object-level training data come from single-session captures and renderings of synthetic 3D assets. They are likewise evaluated on benchmarks built with such data. Two recent works, MegaDepth-X [23] and MegaUnScene [65] train and evaluate feedforward models on in-the-wild internet photo collections, but only on scene-level, not object-centric, collections. TRAINSPOTTING is an object-level dataset similarly built from internet photos. We fine-tune VGGT-Ω on this dataset and compare the result with recent feedforward models on TRAINSPOTTING and NAVI in-the-wild (Section 5).

3. The TRAINSPOTTING Dataset

TRAINSPOTTING is a dataset of objects reconstructed from Wikimedia Commons photos taken by many people over many years. It contains many types of rigid objects (Table 2), including statues, cannons, pipe organs, telephone boxes, fountains, cameras, ships, aircraft, and locomotives. This section describes how our dataset is curated from Internet sources, and Section 4

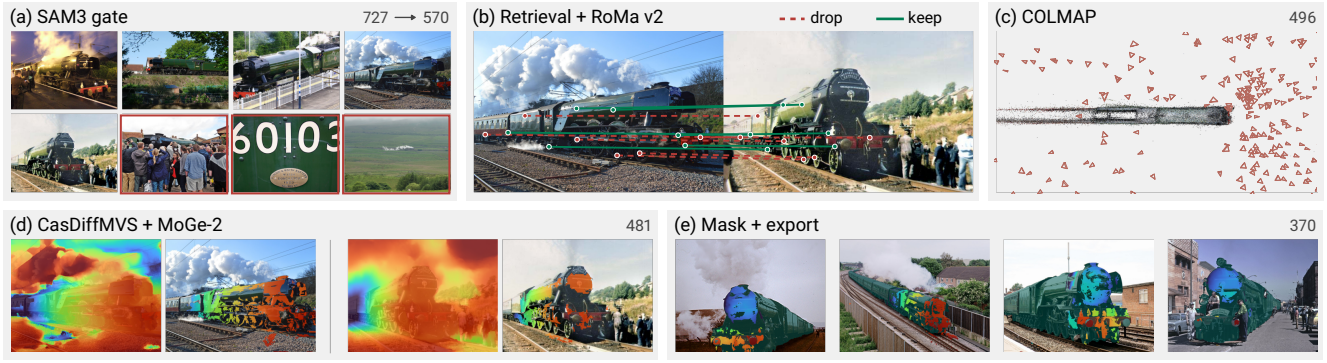


Figure 4. **Reconstruction pipeline for the Flying Scotsman.** (a) Object gating retains 570 of 727 photos, with red borders marking excluded photos. (b) A retrieved pair is shown with candidate correspondences. Green correspondences pass the confidence, cycle, and epipolar checks, while dashed red correspondences fail. (c) The structure-from-motion reconstruction registers 496 images, merged from 14 partial reconstructions. View selection keeps 481 of them for multi-view stereo. (d) The same pair shows MVS depth before and after MoGe-2 filtering. (e) Four of the 370 photos in the dataset show MVS depth in color and the selected mask in teal over the image.

describes how we reconstruct the individual objects.

3.1. Objects from Wikimedia Commons

Wikimedia Commons organizes photos into user-curated categories. Categories for physical subjects are often specific to one object or design, such as *Category:LNER Class A3 4472 Flying Scotsman*. Each object in TRAINSPOTTING corresponds to one such category, and we call the images filed under it a *collection*. Many of these categories are linked from Wikidata, the structured knowledge base supporting Wikimedia projects, and 68% of the TRAINSPOTTING objects have a Wikidata item.

Objects represent either individuals or designs. An *individual* is a single physical unit, such as one hull, one statue, or one numbered locomotive. A *design* is a shape shared by many physical units, such as a car model or a mass-produced telephone box with the same underlying geometry. We pool all units of a design into one single collection, where individual units within the collection may differ in color, trim, and wear.

3.2. Curation at scale

We enumerate candidate categories in two ways. Some classes, such as ships and locomotives, are enumerated from their Wikidata items; others are found by traversing the Wikimedia Commons category graph from root categories, such as *Cameras* and *Pipe organs*. The traversal records 155,744 categories after excluding branches that do not depict the object itself, such as photos taken with a camera model, drawings, and replicas of an object.

The traversal also reaches subcategories for places, events, and parts related to an object, which we wish to exclude. Title rules cannot reliably separate these categories from the objects themselves. For example, *Cannon Street station* is a railway station rather than a cannon. Therefore, we use an LLM to classify each of the recorded categories from its title and the categories above it in the category graph. The classifier assigns three labels: (i) whether the category depicts a rigid physical object, which decides whether it is kept; (ii) whether it names an individual or a design; and (iii) whether the object is permanently installed at one site. Labels (ii) and (iii) determine the regime of each object (Section 3.3). Together with the Wikidata-enumerated classes, the categories kept by the classifier yield 33,884 candidate objects, of which we attempt to reconstruct 17,470 after filtering on the number of photos.

Table 2. **Dataset composition.** Number of objects in each supercategory. An LLM groups the object labels assigned during curation into 53 categories and 13 supercategories.

Supercategory	Objects	Supercategory	Objects
Road vehicles	3,341	Household & personal objects	353
Statues & sculpture	2,160	Weapons & armour	342
Monuments & landmarks	1,773	Religious objects	140
Aircraft & spacecraft	1,461	Street furniture	130
Rail vehicles	1,108	Machinery & equipment	121
Watercraft	701	Other	10
Musical instruments	428	**Total**	**12,068**

This curation alone does not filter individual images by their visual content. Collections retain cluttered photos, partial views, and images in which the object occupies a small part of the frame. We rely on the robust downstream reconstruction pipeline to either register these images in 3D or remove them when they cannot be registered correctly (Section 4).

3.3. Capture variation

Different kinds of objects are characterized by different forms of variation in context and capture. A statue fixed to its plinth yields photos that share a common environment, but differ in viewpoint, time, and lighting. Photos of a locomotive span different stations with corresponding changes in the environment. A telephone-box design represents many physical units whose exact wear and paint may differ. We denote these distinct categories as *fixed*, *roaming*, and *instance-pooled* (Figure 3). *Fixed* denotes an individual installed at one site, *roaming* an individual that moves between sites, and *instance-pooled* a design shared by many physical instances. The labels follow from

LLM classification (Section 3.2) and are therefore approximate.

4. Pseudo-Ground Truth via Reconstruction

A Wikimedia Commons category can contain photos of other objects of the same kind, unrelated subjects, and crowded scenes in which the object is small. In this section, we describe how we robustly reconstruct each object to obtain camera poses, MVS depth, and object masks via object gating, learned dense matching, and geometric verification. Figure 4 follows one object through the reconstruction pipeline.

4.1. Matching and structure from motion

Object gating. We use an LLM to assign a text prompt to each object, similar to the classification in Section 3.2. We then use SAM3 [6] with this prompt to identify images in which the object is the main subject. We keep an image when the largest detected box covers at least 10% of the image. Common rejected images include crowd photos, close-ups of details, and images of a different subject filed in the same category.

Dense matching. We compute point matches between photos with the dense matcher RoMa v2 [16] due to its robustness to environment and object variation. In the supplement, we compare this choice with structure from motion from SIFT, SuperPoint with LightGlue, and MASt3R matches on the same images. We embed each image with the retrieval model of MASt3R-SfM [13] and match it against the 20 closest images in the embedding space. For each image pair, RoMa v2 returns a dense mapping of every pixel of one image to the other along with a certainty map. We sample candidate correspondences from these outputs. We keep a correspondence between a point $\boldsymbol{x}$ in one image and $W(\boldsymbol{x})$ in the other only if

$$\sigma(\boldsymbol{x}) \geq \tau_\sigma, \qquad \left\| W'\big(W(\boldsymbol{x})\big) - \boldsymbol{x} \right\| \leq \tau_{\text{cyc}}, \qquad d\big(W(\boldsymbol{x}), \boldsymbol{F}\boldsymbol{x}\big) \leq \tau_{\text{epi}}, \tag{1}$$

where W and W' are the forward and backward warps, σ is the certainty that RoMa v2 assigns to the match in both directions, and $\boldsymbol{F}$ is a fundamental matrix estimated for the pair with MAGSAC++ [5]. The three conditions only retain confident, cycle-consistent matches that agree with the estimated epipolar geometry. We merge nearby endpoints of the remaining correspondences into shared keypoints, following prior work on structure from motion from dense matches [17, 18, 21]. These shared keypoints link pairwise matches into multi-view tracks.

Structure from motion. We import these keypoints and matches into COLMAP [35] and perform incremental mapping. We exhaustively merge reconstructed components and keep the largest result as the final reconstruction of the object.

4.2. MVS depth and masks

View selection. A mis-registered view, or a view of a different object weakly linked to the reconstruction, is unreliable as ground truth and can corrupt the depth of its neighbors during MVS. Therefore, we keep a view only if enough of its points have a track length of at least three. We then build a covisibility graph in which two views are connected by an edge weighted by the number of 3D points they share. We split this graph by spectral bisection and discard the smaller subgraph when the conductance of the cut is below 0.15.

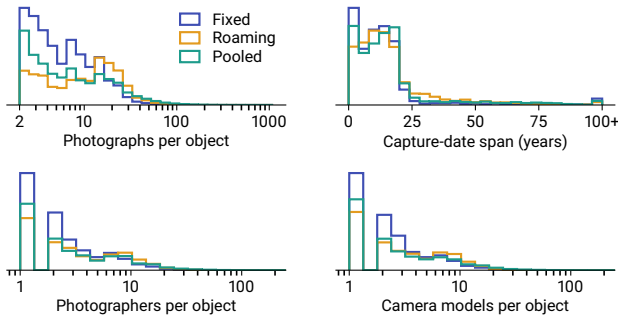


Figure 5. **Object statistics by regime.** Distributions over TRAINSPOTTING of the number of photos, the span of capture dates, the number of photographers, and the number of camera models.

Multi-view stereo with a monocular check. We run Cas-DiffMVS [44] on the selected views and keep high-confidence depths. Since MVS can be unreliable on textureless and specular surfaces, we verify these depths against a monocular depth prediction from MoGe-2 [50], as in MegaDepth-X [23], using the spherical filtering strategy of MoGe-2 and ViGeo [50, 63].

Object masks. We select the SAM3 mask that covers the largest share of the verified depth as the target object mask. A photo is included only if enough of the selected mask carries verified depth. A final quality check removes objects with erroneous reconstructions and photos with implausible focal lengths or in which the object is not the dominant subject.

4.3. Statistics

We pass 17,470 objects with 1,251,986 photos through our reconstruction pipeline. We report the number of objects and photos that remain after each stage in the supplement. Object gating removes the largest share of photos. Structure from motion succeeds for 85% of the objects and registers 72% of the gated photos.

Dataset. Our final dataset, TRAINSPOTTING, contains 12,068 objects and 311,886 photos. Out of all of its objects, 4,338 are fixed, 2,869 roaming, and 4,861 instance-pooled. Capture dates are available for 84% of the photos. The median capture span is 12 years and 90% of the objects have a capture span of at least one year. A median object has photos by 7 different photographers taken with 6 different camera models. Figure 5 shows these distributions by regime. For a coarse view of the composition, we use an LLM to group the 337 object labels of the dataset into 53 categories and 13 supercategories (Table 2). We release TRAINSPOTTING with its camera poses, MVS depth, and object masks. Every photo is distributed with its



Figure 6. **Qualitative comparison.** We visualize ground truth and predicted camera poses and point maps from π^3 [52], Depth Anything 3 [25], VGGT-Ω [49], and SPOTTER on held out objects from TRAINSPOTTING and NAVI in-the-wild [19]. We use 4 or 16 input views for TRAINSPOTTING and 8 for NAVI in-the-wild. Only the first 4 input views are shown. Triangular markers indicate the directions of the cameras outside of the frame. The baselines misregister the input views and produce duplicated or fragmented geometry, whereas SPOTTER recovers a single consistent object and accurate camera poses.

Wikimedia Commons source, author, and licence.

Split. We partition TRAINSPOTTING by object into a training set of 11,068 objects and 279,319 photos and a held-out set of 1,000 objects and 32,567 photos, so that no photo of a held-out object is seen during training.

5. Experiments

5.1. Setup

Model and training. We obtain SPOTTER by fine-tuning the released 1B-parameter checkpoint of VGGT-Ω [49] on the training objects of TRAINSPOTTING. We update the camera head and all attention blocks of the transformer backbone. Each training sample contains one object and between 2 and 24 of its photos, chosen so that they form a connected set under covisibility. We consider two photos covisible when at least 5% of the object points of one project inside the object mask of the other.

Objective. Following VGGT, we express the cameras and depths of each sample in the coordinate frame of its first view. We normalize the ground truth by the mean distance of its valid 3D points to the origin. We normalize the predictions in the same way, following DUSt3R [51]. The model is trained with a combined camera, depth, and point loss:

$$\mathcal{L} = 5\mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{point}}, \tag{2}$$

$$\mathcal{L}_{\text{cam}} = \sum_i \left\| \hat{\boldsymbol{g}}_i - \boldsymbol{g}_i \right\|_1, \tag{3}$$

$$\mathcal{L}_{\text{depth}} = \frac{1}{|\mathcal{V}|} \sum_{\boldsymbol{x} \in \mathcal{V}} \left(1 + \frac{1}{D(\boldsymbol{x})}\right) \left| \hat{D}(\boldsymbol{x}) - D(\boldsymbol{x}) \right|, \tag{4}$$

$$\mathcal{L}_{\text{point}} = \frac{1}{|\mathcal{V}|} \sum_{\boldsymbol{x} \in \mathcal{V}} \left\| \hat{\boldsymbol{p}}(\boldsymbol{x}) - \boldsymbol{p}(\boldsymbol{x}) \right\|_2, \tag{5}$$

where $\hat{\boldsymbol{g}}_i$ is the predicted camera of view i, encoded as translation, rotation quaternion, and field of view, $\mathcal{V}$ is the set of valid pixels, and $\hat{\boldsymbol{p}}(\boldsymbol{x})$ is the 3D point of pixel $\boldsymbol{x}$ given by the predicted depth and camera. For the depth and point losses, we use plain regression loss without confidence weighting and discard the largest 2% of residuals in each batch.

Replay. To limit forgetting, a fixed 15% of every batch

contains samples from BlendedMVS [60], Hypersim [32], and uCO3D [28], supervised with the same loss.

Object masks. Since the background differs between the photos of an in-the-wild collection, the model also predicts an object mask for each view, with a small head trained on the dataset masks on top of the frozen features. This allows restricting the point map output to the object, separating it from its background. The head is trained in a separate stage with the rest of the model frozen and does not affect any of the evaluations below. Details are described in the supplement.

Implementation details. We train with AdamW [29] at a peak learning rate of 10^{-5} and a weight decay of 0.05. We use linear warm-up over the first 5% of steps and a cosine decay. Training runs for 7,500 steps, with 768 images in each step, and takes about 12 hours on four NVIDIA B200 GPUs.

Baselines. We compare SPOTTER with the released checkpoints of VGGT [48], π^3 [52], Depth Anything 3 [25], MapAnything [20], and VGGT-Ω [49].

5.2. Evaluation protocol

Samples. On TRAINSPOTTING, we draw one connected 4-view sample from each held-out object and, when the object has enough photos, one connected 16-view sample, for 1,635 samples from 1,000 objects. On each external dataset, we use one sample of 8 views per scene, equally spaced along the capture sequence when one exists and drawn at random otherwise, since published protocols differ in the number and selection of input views. Samples are drawn with a fixed seed.

Camera pose. We compare the predicted and ground truth cameras on every pair of views in a sample. For each pair, we compute the error of the relative rotation and the direction of the relative translation and report the area under the accuracy curve of the larger of the two errors up to 5° and 30° (AUC@5 and AUC@30), following PoseDiffusion [46]. All metrics are computed per sample and then averaged over samples.

Depth. We evaluate the predicted depth maps on the pixels with ground-truth depth. For models that only predict point maps, we take depth as the z-coordinate of the point map in each camera's frame. We scale the predicted depth of all views in a sample by a single median ratio to the ground truth. We report the absolute relative error (AbsRel) and the inlier ratio $\delta < 1.03$, the fraction of pixels whose predicted and ground-truth depths differ by less than a factor of 1.03.

Point maps. For models that only predict depth, we obtain a point map for each view by unprojecting the predicted depth with the predicted camera. For models that predict point maps, we use the predicted point maps directly. We align the predicted points to the ground-truth points with a similarity transform estimated from corresponding pixels and refined by iterative closest point (ICP). We then report accuracy, the mean distance from each predicted point to the nearest ground-truth point, and completeness, the mean distance from each ground-truth point to the nearest predicted point, following Spann3R [45]. On TRAINSPOTTING, whose reconstructions have an arbitrary scale, we express both in percent of the object extent; on the external datasets, we express them in millimeters.

Table 3. **Results on TRAINSPOTTING.** Point-map accuracy and completeness are given as a percentage of the object extent after a similarity alignment refined by ICP. Depth AbsRel and inlier ratio $\delta_{1.03}$ are measured after a single median scale per sample. Camera-pose AUC at 5° and 30° is computed over all pairs of views (Section 5.2).

Method	Point map		Depth		Camera	
	Acc↓	Comp↓	AbsRel↓	$\delta_{1.03}$↑	AUC_5↑	AUC_{30}↑
VGGT [48]	2.38	8.08	0.160	32.4	11.5	44.1
π^3 [52]	2.33	6.22	0.162	34.7	12.5	46.4
DA3 [25]	2.22	12.17	0.189	32.1	10.1	29.4
MapAny. [20]	2.20	7.75	0.206	26.6	5.3	33.1
VGGT-Ω [49]	1.84	12.68	0.201	28.1	7.2	25.0
SPOTTER	**1.22**	**1.15**	**0.071**	**53.3**	**44.2**	**81.2**

Table 4. **Results by view count and regime.** AUC@30 on the held-out TRAINSPOTTING objects. Each TRAINSPOTTING object contributes one 4-view sample and, if it has at least 16 photos, one 16-view sample. The regime columns pool both sample sizes.

Method	4 views (1000)	16 views (635)	Fixed (442)	Roaming (450)	Pooled (743)
VGGT-Ω [49]	29.4	18.1	56.7	14.9	12.3
SPOTTER	**79.9**	**83.2**	**79.0**	**81.0**	**82.6**

5.3. Results on TRAINSPOTTING

Table 3 reports evaluation results on TRAINSPOTTING. SPOTTER improves on VGGT-Ω in every metric, most strongly in camera pose and completeness, and substantially outperforms the strongest baseline, π^3. Table 4 shows that the difficulty is concentrated in roaming and instance-pooled objects, whose photos differ most in environment. VGGT-Ω performs considerably better on fixed objects than on the other two regimes, whereas SPOTTER performs similarly on all three, and performance further improves with more views. Figure 6 shows reconstructions of four TRAINSPOTTING objects and two NAVI in-the-wild objects. The baselines fail to register views taken in different environments. Their point maps therefore contain duplicated or fragmented copies of the object, and their cameras are scattered. SPOTTER registers the same views consistently and reconstructs a single object.

5.4. NAVI in-the-wild

The NAVI dataset [19] contains 36 objects, each with a scanned mesh and camera poses aligned to it, photographed both in *multiview* captures recorded in a single session and in an *in-the-wild* collection across different environments. As the held-out TRAINSPOTTING results are scored against pseudo-ground truth from our own pipeline, evaluating on NAVI in-the-wild provides an independent measure of accuracy. Under our protocol NAVI yields 35 in-the-wild and 281 multiview samples. Tables 5 and 6 report the results. VGGT-Ω reaches an AUC@30 of 91.6 on the

Table 5. **Geometry on external benchmarks.** Point-map accuracy and completeness are given in millimeters after a similarity alignment refined by ICP. Depth inlier ratio $\delta_{1.03}$ is measured after a single median scale per sample. Every model is given the same 8 views of each scene. NAVI photographs each object in one session (*NAVI*) and across different environments (*NAVI (wild)*). The other datasets record each scene in a single capture.

Method	NAVI (wild)			NAVI			7-Scenes			NRGBD			DTU			ETH3D		
	Acc↓	Comp↓	$\delta_{1.03}$↑	Acc↓	Comp↓	$\delta_{1.03}$↑	Acc↓	Comp↓	$\delta_{1.03}$↑	Acc↓	Comp↓	$\delta_{1.03}$↑	Acc↓	Comp↓	$\delta_{1.03}$↑	Acc↓	Comp↓	$\delta_{1.03}$↑
VGGT [48]	3.2	11.5	35.0	2.5	3.4	71.5	25.3	32.6	67.5	17.4	16.9	90.9	**1.2**	1.8	**97.4**	291.4	317.5	52.6
π^3 [52]	3.7	9.6	41.2	**2.2**	**3.0**	76.5	**20.0**	**26.7**	70.5	17.9	17.2	88.6	**1.2**	**1.5**	91.2	154.7	129.7	62.4
DA3 [25]	3.3	16.0	41.1	3.3	5.7	**80.0**	25.4	35.7	74.1	16.4	15.3	**96.7**	1.7	1.6	96.5	274.3	270.3	73.2
MapAny. [20]	4.1	11.0	29.6	3.8	5.8	56.4	28.2	36.8	58.3	42.8	30.2	69.2	2.4	1.9	37.8	459.7	542.0	51.9
VGGT-Ω [49]	**3.1**	20.4	29.7	2.9	5.3	76.1	21.8	27.4	71.9	13.3	9.0	92.4	**1.2**	1.7	97.1	73.6	61.1	**83.5**
SPOTTER	3.2	**1.9**	**43.1**	3.1	4.0	74.5	20.2	27.4	**75.1**	**12.9**	**8.3**	92.7	**1.2**	1.6	95.9	**65.5**	**52.9**	82.8

Table 6. **Camera pose on external benchmarks.** Camera-pose AUC at 5° and 30° is computed over all pairs of the same 8 views of each scene.

Method	NAVI (wild)		NAVI		7-Scenes		NRGBD		DTU		ETH3D	
	AUC@5↑	AUC@30↑	AUC@5↑	AUC@30↑	AUC@5↑	AUC@30↑	AUC@5↑	AUC@30↑	AUC@5↑	AUC@30↑	AUC@5↑	AUC@30↑
VGGT [48]	14.4	45.5	53.3	91.2	26.6	81.0	86.4	97.6	**98.1**	**99.7**	37.6	71.1
π^3 [52]	13.1	43.7	55.8	91.6	28.8	83.3	86.5	97.2	80.0	96.6	36.4	72.9
DA3 [25]	12.0	36.9	**57.2**	91.8	28.5	81.5	**91.4**	**98.6**	91.0	98.5	55.4	75.4
MapAny. [20]	3.3	24.2	28.4	84.0	18.2	72.4	56.5	90.9	15.0	73.4	33.8	65.6
VGGT-Ω [49]	3.8	18.8	55.4	91.6	**40.8**	**87.2**	87.8	97.9	91.1	98.5	**66.9**	**91.8**
SPOTTER	**30.7**	**76.0**	56.7	**91.9**	38.8	86.1	86.5	97.3	86.1	97.7	65.4	**91.8**

multiview captures but only 18.8 on the in-the-wild collections of the same objects. Fine-tuning on TRAINSPOTTING raises the in-the-wild AUC@30 to 76.0 and improves completeness from 20.4 to 1.9 mm. SPOTTER achieves the best in-the-wild result on every metric except accuracy, on which all models lie within 1 mm of one another. We note that when the views of a prediction are misregistered, the similarity alignment scales the prediction down toward the center of the object, so each predicted point stays close to the surface while most of the surface is left uncovered. Pose failures therefore appear as completeness error rather than accuracy error. On the multiview captures, performance is largely unchanged.

5.5. Conventional benchmarks

We evaluate all models on 7-Scenes [37], Neural RGB-D [4], DTU [1], and ETH3D [36] to measure the effect of fine-tuning on conventional data (Tables 5 and 6). Relative to VGGT-Ω, fine-tuning changes AUC@30 by at most 1.1 points and $\delta < 1.03$ by at most 3.2 points on these four datasets.

5.6. Ablations

Table 7 compares models trained on subsets of the TRAINSPOTTING training objects. All models in this comparison, including the reference model trained on all objects, are trained for 10 epochs with the same recipe as SPOTTER. We find that the types of photos used for training matter more than their scale. Training only on fixed objects yields lower accuracy than training on a random subset half its size, both on held-out objects and on NAVI in-the-wild. It is also considerably less accurate on roaming and instance-pooled objects, whereas models trained on those regimes remain close to the reference model when evaluated on fixed objects. This shows that photos across different environments, which fixed objects and existing datasets largely lack, are necessary for generalization to the other regimes and to NAVI in-the-wild.

Table 7. **Training-data ablations.** AUC@30 of models trained for 10 epochs on subsets of the TRAINSPOTTING training objects. The number of training objects is given in parentheses.

	TRAINSPOTTING held-out				
Training data	All	Fixed	Roaming	Pooled	NAVI (wild)
All (11,068)	80.3	78.2	79.5	82.1	74.6
Random 20% (2,193)	79.7	77.5	79.2	81.3	72.3
Fixed (4,038)	68.5	78.4	58.0	68.9	63.4
Roaming (2,603)	79.0	76.5	80.3	79.7	65.7
Pooled (4,427)	78.8	75.6	77.0	81.7	73.9

6. Conclusion

Current feedforward reconstruction models perform poorly on in-the-wild object-centric photo collections, even for objects that they reconstruct accurately from a single capture session. Such in-the-wild collections are abundant in public photo archives, and TRAINSPOTTING shows that they can be turned into training data at scale. Fine-tuning VGGT-Ω on TRAINSPOTTING improves camera pose, depth, and 3D geometry on held-out objects and on the independent NAVI in-the-wild benchmark, while performance on conventional benchmarks remains largely unchanged. The base models had already been trained on large datasets of objects recorded in single capture sessions, and we also show that training on fixed objects alone transfers poorly to roaming and instance-pooled objects, so what these models lacked was variation in how an object is photographed. Beyond training, TRAINSPOTTING provides a benchmark for reconstruction under changing environments, and its reconstructions of objects photographed over decades allow studying how objects and their surroundings change over time.

Acknowledgments

This material is based upon work supported by the Naval Air Warfare Center under Contract No. N68335-25-C-0482. We gratefully acknowledge use of the research computing resources of the Empire AI Consortium, Inc, with support from Empire State Development of the State of New York, the Simons Foundation, and the Secunda Family Foundation.

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *IJCV*, 120:153–168, 2016. 2, 8

[2] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *ICCV*, 2009. 3

[3] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 12

[4] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *CVPR*, 2022. 8

[5] Daniel Barath, Jana Noskova, Maksym Ivashechkin, and Jiří Matas. MAGSAC++, a fast, reliable and accurate robust estimator. In *CVPR*, 2020. 5

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026. 5

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2

[9] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-XL: A universe of 10m+ 3D objects. In *NeurIPS Datasets and Benchmarks*, 2023. 2

[10] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2

[11] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Workshops*, 2018. 3, 12

[12] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A high-quality dataset of 3D scanned household items. In *ICRA*, 2022. 2

[13] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025. 3, 5, 12

[14] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023. 3

[15] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, 2024. 3

[16] Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. RoMa v2: Harder better faster denser feature matching, 2025. arXiv:2511.15706. 3, 5

[17] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-free one-shot object pose estimation without CAD models. In *NeurIPS*, 2022. 5

[18] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *CVPR*, 2024. 3, 5

[19] Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin-Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-agnostic image collections with high-quality 3D shape and pose annotations. In *NeurIPS Datasets and Benchmarks*, 2023. 2, 6, 7

[20] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *3DV*, 2026. 3, 7, 8, 19

[21] JongMin Lee and Sungjoo Yoo. Dense-SfM: Structure from motion with dense consistent matching. In *CVPR*, 2025. 3, 5

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

[23] Yuan Li, Yuanbo Xiangli, Hadar Averbuch-Elor, Noah Snavely, and Ruojin Cai. Long-tail internet photo reconstruction. In *CVPR*, 2026. 2, 3, 5

[24] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2, 3

[25] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026. 3, 6, 7, 8, 19

[26] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *ICCV*, 2023. 12

[27] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao

Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. DL3DV-10K: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, 2024. 2

[28] Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y. Zhang, Natalia Neverova, Andrea Vedaldi, Roman Shapovalov, and David Novotny. UnCommon objects in 3D. In *CVPR*, 2025. 1, 2, 7

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[30] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 12

[31] 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 *ICCV*, 2021. 1, 2

[32] 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 *ICCV*, 2021. 7

[33] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 12

[34] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3

[35] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 3, 5, 12

[36] Thomas Schöps, Johannes L Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 8

[37] 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 *CVPR*, 2013. 8

[38] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics*, 25(3):835–846, 2006. 2, 3

[39] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *CVPR*, 2021. 3

[40] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3D: Dataset and methods for single-image 3D shape modeling. In *CVPR*, 2018. 2

[41] 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 *CVPR*, 2025. 3

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

[43] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 13

[44] Fangjinhua Wang, Qingshan Xu, Yew-Soon Ong, and Marc Pollefeys. Lightweight and accurate multi-view stereo with confidence-aware diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11):10060–10075, 2025. 5

[45] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *3DV*, 2025. 7

[46] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *ICCV*, 2023. 7

[47] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *CVPR*, 2024. 3

[48] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025. 1, 3, 7, 8, 19

[49] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *CVPR*, 2026. arXiv:2605.15195. 1, 3, 6, 7, 8, 13, 19

[50] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025. 5

[51] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024. 3, 6

[52] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, 2026. 1, 3, 6, 7, 8, 19

[53] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. OmniObject3D: Large-vocabulary 3D object dataset for realistic perception, reconstruction and generation. In *CVPR*, 2023. 2

[54] Yushuang Wu, Luyue Shi, Haolin Liu, Hongjie Liao, Lingteng Qiu, Weihao Yuan, Xiaodong Gu, Zilong Dong, Shuguang Cui, and Xiaoguang Han. MVImgNet2.0: A larger-scale dataset of multi-view images. *ACM Transactions on Graphics*, 43(6): 173:1–173:16, 2024. 2

[55] 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 *CVPR*, 2024. 1, 2

[56] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *WACV*, 2014. 2

[57] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. ObjectNet3D: A large scale database for 3D object recognition. In *ECCV*, 2016. 2

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

[59] 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 *CVPR*, 2025. 3

[60] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020. 7

[61] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 2

[62] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. MVImgNet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 2

[63] Zhu Yu, Jingnan Gao, Runmin Zhang, Lingteng Qiu, Zhengyi Zhao, Rui Peng, Yichao Yan, Kejie Qiu, Siyu Zhu, Zilong Dong, Si-Yuan Cao, and Hui-Liang Shen. Towards consistent video geometry estimation. *arXiv preprint arXiv:2605.30060*, 2026. 5

[64] 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 *CVPR*, 2025. 3

[65] Yiwen Zhang, Joseph Tung, Ruojin Cai, David Fouhey, and Hadar Averbuch-Elor. Emergent extreme-view geometry in 3D foundation models. In *CVPR*, 2026. 3

Supplementary Material

A. Dataset Curation

This section gives the details of the curation summarized in Section 3.2 of the main paper.

Enumeration. Thirteen vehicle classes are enumerated from their Wikidata items: aircraft, ships, locomotives, automobiles, rail vehicles, buses, trams, military vehicles, trucks, motorcycles, tractors, construction equipment, and other vehicles. Each item gives one Wikimedia Commons category, 22,983 in total. Whether a class names individual objects or designs is fixed by the class: aircraft, ships, locomotives, and trams are identified by registration, hull, or fleet number, while the remaining classes are identified by model.

All other objects are found by traversing the Wikimedia Commons category graph from 856 root categories. We chose 211 of them by hand, and the LLM screened the remaining 645 from the subcategories of broad hand-chosen categories. The traversal does not descend into branches whose titles match one of seven kinds of category that do not depict the object itself: photos taken with a camera model, people with an object, media about an object, replicas and scale models, institutions, parts, and wiki maintenance categories. It prunes 37,930 branches this way and records 155,744 categories.

Classification. Of the recorded categories, 49,252 contain at least 25 photos, counting the photos of subcategories that depict the same object. The LLM classifies each of them from its title and the titles of the categories above it in the graph. The prompt is reproduced in Section G. It keeps 40,254 categories as rigid physical objects, rejects 8,956 for the reasons listed in Table S2, and leaves 42 undecided. Of the kept categories, 13,347 name an individual object or a design, and the remaining 26,907 are categories of many objects, such as *Statues in Paris*, which we do not use. After removing categories already covered by the vehicle classes and merging subcategories that depict the same object into their parent, 10,901 object categories remain. Together with the vehicle classes, they form the 33,884 candidate objects of the main paper.

Table S2. **Categories rejected by the classifier.** Reasons for the 8,956 categories with at least 25 photos that the LLM does not keep.

Reason	Categories
Flat object, such as a painting or a sign	1,802
Media or two-dimensional depiction of an object	1,606
Place or institution	1,489
Abstract or maintenance category	1,066
Event or activity	811
Deformable object	767
Part of an object	614
People	275
Photos taken with or of a product	246
Other	280
Total	8,956

Categories. Table S1 lists the 53 categories into which the LLM groups the 337 object labels, with the number of objects and photos of each.

Overlap with internet-photo scene datasets. Table S3 gives the overlap of Section 2 of the main paper by regime. A category counts as a MegaScenes scene when its title matches a category collected by MegaScenes, and as reconstructed when MegaScenes releases a reconstruction for it. Nearly all overlapping objects are fixed objects, which MegaScenes and MegaDepth-X treat as scenes.

Table S3. **Overlap with internet-photo scene datasets by regime.** Number of TRAINSPOTTING objects whose Wikimedia Commons category is a MegaScenes scene, has a released MegaScenes reconstruction, or is a MegaDepth-X scene.

Regime	Objects	MegaScenes category	MegaScenes reconstruction	MegaDepth-X
Fixed	4,338	1,798	1,128	56
Roaming	2,869	41	34	0
Instance-pooled	4,861	2	1	0
Total	12,068	1,841	1,163	56

B. Reconstruction Pipeline

Table S4 lists the objects and photos that remain after each stage of the pipeline. Table S5 lists the fixed thresholds of the pipeline of Section 4 of the main paper.

Comparison with other matchers. Figure S1 compares the pipeline of Section 4 of the main paper with four alternatives on the images that pass object gating: COLMAP with SIFT and exhaustive matching [30, 35], SuperPoint with LightGlue [11, 26] in hloc [33] with NetVLAD retrieval [3], MASt3R matches [22] with the COLMAP mapper, and MASt3R-SfM [13]. Each alternative runs with its default settings and, where it retrieves image pairs, with 20 neighbours per image as in our pipeline. On the instance-pooled and roaming objects, the alternatives fail to reconstruct the object. Even on the fixed objects, our pipeline produces a more complete and cleaner reconstruction.

C. Training Details

This section completes the setup of Section 5.1 of the main paper.

Trainable parameters. We fine-tune the released VGGT-Ω checkpoint with 1B parameters. All 24 frame and global attention layers of the aggregator and the camera head are trained, while the patch embedding, the camera and register tokens, and the depth head are frozen. This leaves 808M of the 1.14B parameters trainable.

Sampling. The number of views per sample is drawn once per batch, uniformly between 2 and 24. Each batch is filled

Table S1. **Object categories.** The 53 categories, grouped by supercategory, with the number of objects and photos of each (12,068 objects and 311,886 photos in total).

Category	Objects	Photos	Category	Objects	Photos
Road vehicles			*Musical instruments*		
car	1,967	87,193	pipe organ	390	5,733
bus	476	28,932	musical instrument	38	982
military vehicle	341	7,522	*Household & personal objects*		
truck & van	269	8,553	camera & lens	142	2,946
motorcycle	187	2,776	computer & office machine	90	1,364
tractor & other road vehicle	71	522	vase & vessel	47	893
bicycle	30	506	consumer electronics	33	457
Statues & sculpture			appliance & personal item	24	260
statue	1,628	24,940	furniture & lighting	17	178
sculpture	432	7,649	*Weapons & armour*		
bust & head	100	1,576	artillery	173	2,088
Monuments & landmarks			firearm & hand weapon	127	1,550
war memorial	553	8,079	missile	29	286
fountain	385	4,788	helmet	13	144
monument & memorial	340	8,276	*Religious objects*		
obelisk	141	2,205	altar & altarpiece	68	504
gravestone & tomb	76	1,173	church furnishing & reliquary	30	291
megalith	73	1,131	cross & crucifix	25	261
sarcophagus & coffin	46	685	bell	17	261
door, gate & arch	45	949	*Street furniture*		
clock & sundial	39	1,005	post box	85	3,003
runestone & carved stone	33	657	bench & other street furniture	28	325
commemorative column & pillar	31	302	telephone box	17	733
tower & small building	11	182	*Machinery & equipment*		
Aircraft & spacecraft			engine	62	618
aircraft	1,415	37,798	industrial machine & equipment	24	385
spacecraft & rocket	46	647	crane	14	135
Rail vehicles			scientific instrument	12	171
locomotive	543	16,838	robot	9	111
tram	496	16,033	*Other*		
train & railcar	69	2,493	other object	10	61
Watercraft					
ship & boat	701	14,736			

Table S4. **Reconstruction yield.** Number of objects and photos that remain after each stage of the pipeline in Section 4 of the main paper.

Stage	Objects	Photos
Input	17,470	1,251,986
Object gating	16,141	775,609
Structure from motion	14,818	558,281
MVS depth	14,032	486,275
Mask and depth checks	13,236	367,245
Final dataset	12,068	311,886

to a fixed budget of 192 images per GPU. Inputs follow the training preprocessing of VGGT-Ω [49]. Depth and points are supervised only on object pixels that have MVS depth. The replay samples from BlendedMVS, Hypersim, and uCO3D are supervised on all pixels with ground truth.

Optimization. The cosine schedule decays the learning rate to 10^{-8}. Gradients are clipped to a norm of 1. Training uses bfloat16 precision with gradient checkpointing.

Object mask head. After fine-tuning, we freeze the model and train a small convolutional head on the fused features of its depth head to predict, for each pixel, whether it belongs to the object. The target is the object mask of the dataset. On held-out objects, the predicted masks reach an intersection over union of 0.92 with the dataset masks.

D. Evaluation Details

This section details the protocol of Section 5.2 of the main paper.

Samples on TRAINSPOTTING. The views of a sample form a connected set under the covisibility relation used for training: two photos are connected when at least 5% of the object points of one project inside the object mask of the other. Each view is resized to cover 512×512 pixels and cropped around its principal point.

External datasets. For ordered captures, the 8 views are equally spaced over the sequence, including its first and last frame. For unordered collections, they are a fixed random draw. Scenes with fewer than 8 usable frames are skipped. On NAVI, images are center-cropped to a square, and frames that the dataset flags as occluded are excluded. Inputs are resized to each model's patch multiple, and outputs are resampled to the ground-truth grid.

Metrics. The depth scale of a sample is the median ratio of ground-truth to predicted depth over the valid pixels of all its views. The similarity transform of the point-map metrics is estimated from corresponding valid pixels [43] before the ICP refinement.

Effect of the alignment on misregistered predictions. Section 5.4 of the main paper explains why pose failures appear as completeness error rather than accuracy error. Table S6

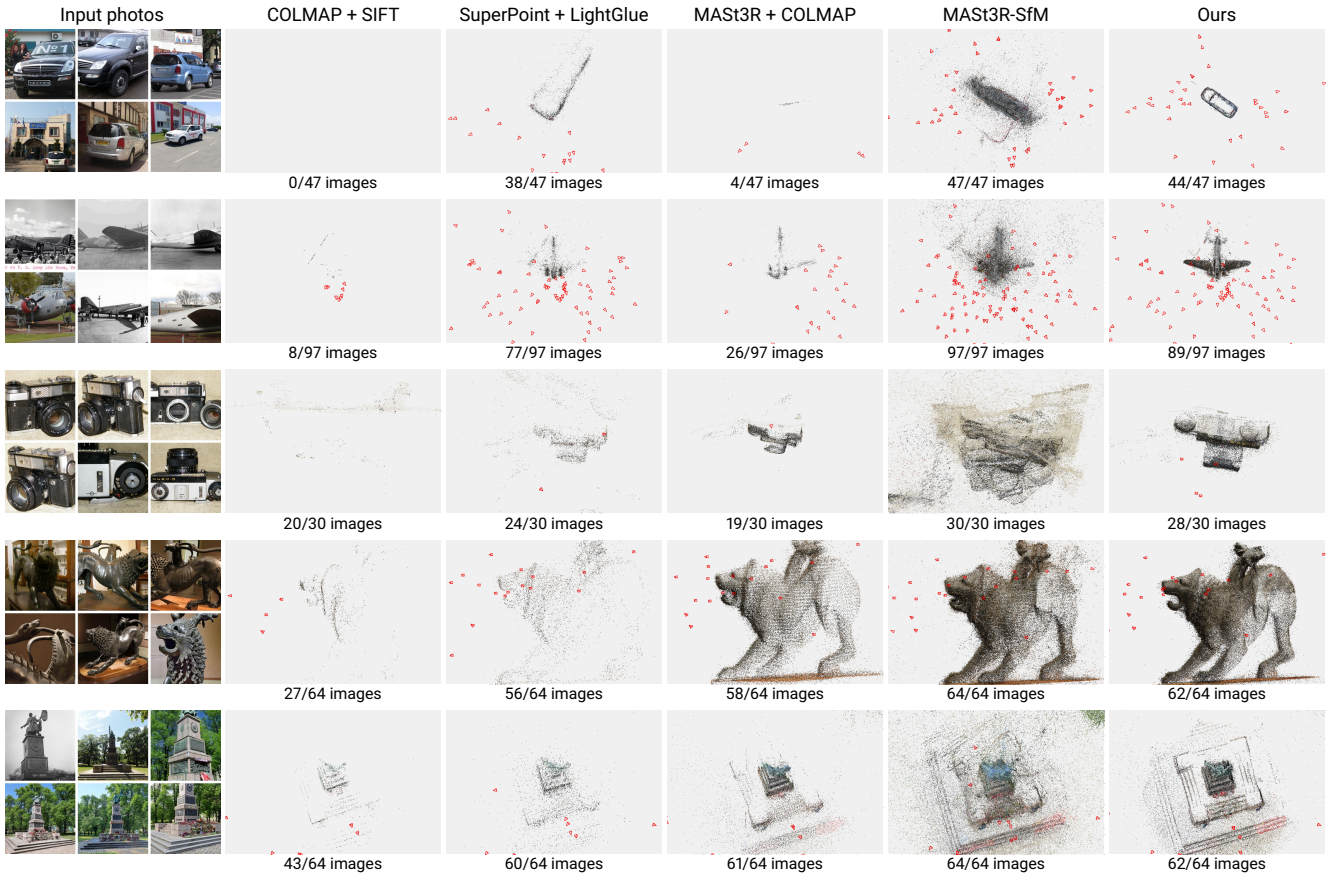


Figure S1. **Structure from motion with other matchers.** Each row shows one TRAINSPOTTING object: six of the input photos, then the largest model of COLMAP with SIFT, SuperPoint with LightGlue, MASt3R matches with the COLMAP mapper, MASt3R-SfM, and our pipeline, all run on the same images. The number of registered images out of the input images is given under each panel. The other methods fail on instance-pooled and roaming objects, and on fixed objects our pipeline produces a more complete and cleaner reconstruction. In the bottom three rows, we frame the object to show its details, so not all cameras are visible.

quantifies this on NAVI in-the-wild by the extent of the aligned prediction relative to the extent of the ground truth. The alignment scales most predictions of VGGT-Ω to less than half the object's size, while the predictions of SPOTTER keep their size.

Table S6. **Extent of the aligned prediction on NAVI in-the-wild.** Ratio of the extent of the aligned prediction to the extent of the ground truth over the 35 objects, and the number of objects with a ratio below 0.7.

Method	Median	10th pct.	90th pct.	Objects below 0.7
VGGT-Ω	0.43	0.15	0.87	28
SPOTTER	0.92	0.81	0.96	0

E. Additional Qualitative Results

Failure cases. Figure S2 shows four held-out samples on which SPOTTER fails, one for each of the most common failure modes. Views taken with long focal lengths are ambiguous between focal length and depth, so their points can be misaligned with the other views and appear as noise around the airliner. Close-up views can be ambiguous in their placement on the object, and the close-up of a wheel is localized at the wrong position on the car. Symmetric objects produce doppelganger ambiguities, and the cameras of the grave monument are placed on the wrong one of its two identical pillars. Objects far from the training distribution, such as the toy shark, whose appearance changes drastically with its pose and placement across the views, are reconstructed inconsistently.

Additional qualitative comparisons. Figure S3 shows ten held-out samples drawn at random.

F. Additional Results

Table S7 reports all metrics of every model on TRAINSPOTTING, and Table S8 on the external benchmarks.

G. LLM Prompts

All three LLM passes of the curation use the same setup. The model is Claude Opus 5, the system prompt is the rubric reproduced below, each user message contains a batch of 50 numbered categories, and the response is JSON constrained

Table S5. **Pipeline thresholds.** The fixed thresholds of the reconstruction pipeline of Section 4 of the main paper, in the order of its stages. Image-relative values are fractions of the image side.

Stage	Threshold	Value	Definition
Object gating	minimum box size	10% of the image area	Largest SAM3 detection, with detection score above 0.5.
Retrieval	neighbours per image	20	
Matching	correspondences per pair	8,000	Sampled from the dense warp.
	τ_σ	0.20	Certainty threshold.
	τ_{cyc}	0.4% of the image side	Cycle-consistency distance.
	τ_{epi}	0.1% of the image side	Epipolar distance.
	keypoint merge radius	0.2% of the image side	Match endpoints within this radius form one keypoint.
View selection	confirmed points	100	Points on tracks of length at least 3.
	confirmed fraction	0.30	Confirmed points over all triangulated points of the image.
	covisibility cut	conductance below 0.15	Spectral bisection of the covisibility graph; the largest cluster is kept.
Multi-view stereo	source views	6	
	confidence	0.3, 0.5, 0.5	Minimum confidence at the three stages of CasDiffMVS.
Mask selection	confident depth	200 px	Minimum confident MVS pixels for a mask to be chosen.
	verified depth	5% of the mask	Minimum fraction of the chosen mask with verified depth.
Final quality check	dominant subject	5×	Chosen mask area relative to every other detection.
	focal length	0.3 to 10 image widths	
	photos per object	2 (training), 5 (held-out)	

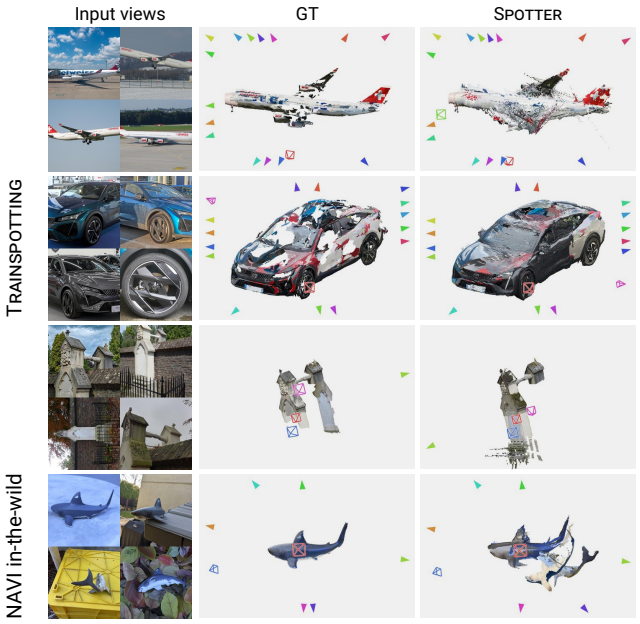


Figure S2. **Failure cases.** Four held-out samples on which SPOTTER fails: airliner (noisy depth at long focal lengths), car (misregistered close-up of a wheel), grave monument (doppelganger ambiguity), and toy shark from NAVI in-the-wild (out of distribution).

to a fixed schema. On a hand-labelled set of 32 categories, the classification pass agreed with the labels on all 32 verdicts and on 14 of 15 granularity labels.

Category classification (Section 3.2 of the main paper). The response for each category gives the verdict (keep, reject, or uncertain), the reject reason, the granularity (individual, model, named unknown, or pool), and whether the object is installed at a fixed site.

```
You are classifying Wikimedia Commons category titles
for a 3D object reconstruction dataset.  Each category's
images will be used to reconstruct a rigid physical
object with COLMAP. Your job is SUBJECT identity:  decide
from the title (plus its parent category and crawl root)
whether the category is about physical objects suitable
for reconstruction.

Each numbered item gives:
- the category TITLE (your main evidence),
- parent:  the category it was found under,
- root:  the top-level category a breadth-first crawl
  started from to reach it.  A hint only -- the crawl
  sometimes wanders, so an item may sit under an unrelated
  root.  An off-root category of genuine physical objects
  is still a keep.
- files:  how many image files the category holds,
- wikidata:  whether a Wikidata item is linked to it.  A
  weak hint that the category names a recognized concept
  -- but concepts include generic types ("settee"), so it
  proves nothing about specificity.

The pipeline tolerates noise but not wrongness:  a re-
jected category is gone forever, while a wrongly kept one
merely wastes a download before reconstruction discards it.
When a title does not clearly establish a reject reason,
KEEP it.

For each numbered item return:

verdict:
- "keep" -- the category depicts physical, rigid objects
  (whatever family they belong to).
- "reject" -- the title alone establishes one of the reject
  reasons below.
- "uncertain" -- genuinely cannot tell what the category
  depicts.

reason (for reject; use "ok" when keep):
- "usage_or_provenance" -- photos taken WITH the object
  or sourced from somewhere:  "Taken with Nikon D5300",
  "Photos from X stream", "Media from iNaturalist".
- "people" -- people using/holding/wearing the object:
  "Guitarists with...", "People with Canon cameras".
- "media_or_2d" -- depictions rather than the object:
  logos, icons, SVG, diagrams, drawings, paintings, stamps,
  advertisements, book titles, close-up detail shots.
- "place_or_institution" -- a place, street, building,
  company, museum, maker, or event venue that merely
  contains the family noun:  "Cannon Street station",
  "Scientific instrument makers".
- "event_or_activity" -- competitions, festivals, cer-
  emonies, activities:  "2019 ICF Canoe Slalom World
  Championships", "Emptying of post boxes".
- "abstract_or_meta" -- licenses, wiki maintenance,
  pronunciation audio, quality-image collections,
  categories-about-categories.
- "flat" -- near-planar subjects:  manhole covers, coins,
  plaques, murals, signage, banknotes.
- "deformable_or_soft" -- clothing, uniforms, textiles,
  bags, plush.
- "parts_or_accessories" -- components rather than whole
  objects:  "Guitar headstocks", "Camera parts", "Bicycle
```

```
   pannier racks" is a borderline whole object (keep); use
   this only for clear fragments of a larger object.
- "other" -- a reject that fits none of the above (say
   nothing else; the verdict is what matters).

granularity (for keep; "n_a" for reject):
- "individual" -- one specific physical object:  "Penfold
   pillar box on Otterndorf Green, Sheringham", "USCGC
   White Pine (WLM-547)".
- "model" -- one manufactured design or standardized type
   whose units are geometrically interchangeable:  "Fender
   Stratocaster", "K6 telephone boxes in Derbyshire",
   "Queen Elizabeth II type A/B red pillar boxes in
   Merseyside", "IBM Model M Keyboard".  A place qualifier
   does not demote a model.
- "named_unknown" -- the title carries a proper-name or
   model-number token you do not recognize, modifying
   the family noun:  "Parc Vue benches in New York City",
   "KX100 telephone boxes".  An unfamiliar name is evidence
   of specificity -- never downgrade these to pool just
   because you don't know the name.
- "pool" -- a generic type, possibly place-scoped, con-
   taining many different designs:  "Snowplows", "Settees",
   "Traffic bollards in Wuppertal", "Cannons in Florida",
   "Wooden benches".  Also under-specified family names
   that span multiple geometries:  "Glock" (17/19/26/43
   differ), "King George V post boxes in the Republic of
   Ireland" (mixes wall-mounted and pillar).

fixed_location (boolean):  true if the objects are
permanently installed at fixed sites (street furni-
ture, installed fountains, monuments-like); false if
portable/movable; judge from the object family when the
title doesn't say.

object_scale (boolean, for keep; use false for reject):  a
descriptive FLAG, never a reason to reject.  True if the
subject is hand-to-room scale -- a statue, gravestone,
cannon, telescope on a mount, church organ.  False if
larger.  The test:  could a person walk around it and
photograph the whole thing from a few metres?  "Very Large
Telescope" (an observatory site) -> false; "Zenit avtomat"
(a camera) -> true.  Large MOVABLE things -- ships,
airliners, locomotives -- are false and that is fine; the
flag records size only, and such categories are still kept.
Obelisks and triumphal columns -> true; anything with
interior rooms -> false.

Different PHYSICAL UNITS of one design count as one object.
A category like "Honda Accord Crosstour" contains photos
of many distinct physical cars -- different owners, colors,
countries -- yet it is ONE object for reconstruction,
because every unit shares the same body shape.  Likewise
fifty K6 telephone kiosks across fifty towns are one
"model".  A category is never a pool merely because its
photos show more than one physical specimen.

The line between "model" and "pool" is SHAPE, not
appearance.  Reconstruction only uses geometry:  units
that differ in color, finish, livery, trim, or wear but
share one body shape are ONE model ("Fender Stratocaster"
stays a model despite countless colors; "Red post boxes"
pooled across designs is still a pool despite one color).
A category is a pool only when the objects in it have
materially different SHAPES -- different products, differ-
ent size variants (Glock 17 vs 26), or a bare type noun
spanning many designs.  Minor sub-variant drift (a changed
headstock logo, a facelifted grille) does not demote a
model; a doubt about shape goes to "model", not "pool".

Rules of thumb:
- Distinguish "type + place" pools from "named design +
   place" models by whether the pre-place part names a
   specific design (code, reign+type, brand+model) or a
   bare noun.
- A brand alone ("Zenit cameras") is a pool of that brand's
   models -- keep, pool.
- Museum/display context does not reject an object
   category:  "Torpedoes" in museums are still torpedoes.
   Only reject when the subject is the institution itself.
- When two readings are plausible, keep it and give your
   best-guess granularity -- do not resolve doubt by
   rejecting.
```

User message:

```
Classify these Wikimedia Commons categories per the
rubric:
0.  "<title>" | parent:  <parent category> | root:
<crawl root> | files:  <n> | wikidata:  yes/no
1.  ...
```

Segmentation prompt (Section 4.1 of the main paper). The response for each category gives one generic noun phrase, optional alternatives, and a confidence.

```
You are naming the object class of Wikimedia Commons
categories.  Each category is about one physical object
(or one manufactured design).  Your output is a single
generic noun phrase per category -- the name of the kind
of thing it is.  It will later be used as a text prompt
for an image segmentation model; you do not need to know
anything about that model.  Just name the object correctly
and consistently.

Each numbered item gives the category TITLE, its parent
category, the crawl root it was reached from, and a
granularity label (individual / model / named_unknown /
pool) from an earlier pass.

For each item return:

sam_prompt -- ONE short generic noun phrase for the object
class:  "statue", "cannon", "gravestone", "pipe organ",
"telephone box", "typewriter", "wind turbine".  Rules:
- Generic, never specific.  "Bust of Voltaire" -> "bust".
   "Pentax K10D" -> "camera".  "K6 telephone boxes in
   Derbyshire" -> "telephone box".  Names, model numbers,
   places, and reigns never appear in the output.
- Name the WHOLE object the category is about, not a
   component:  "pipe organ" not "organ pipes"; "fountain"
   not "basin"; "cannon" not "cannon barrel".
- No material, color, pose, or style:  "bronze equestrian
   statue" -> "statue".  Those describe one instance; the
   same word must fit every unit in the category.
- Two-word compounds are fine when a single word is wrong
   or too broad:  "wayside cross", "baptismal font",
   "sewing machine", "vending machine".

ANCHOR TO THE CATEGORY TREE. The parent and root tell
you the object's family, and that is authoritative:
everything under a "Dolmens" root is a dolmen, everything
under "Pipe organs by country" is a pipe organ.  Use the
family's noun unless the TITLE clearly names something
more specific or different (a "War memorials" root can
hold an obelisk, a statue, or a wall -- take the title's
word there).  Root and parent names are index-shaped
("Coastal artillery", "War memorials by country") -- give
the object noun ("cannon", "war memorial"), not the index
name.  Be CONSISTENT: every category in the same family
should get the same word unless its title says otherwise.

alternatives -- normally an EMPTY list.  Fill it ONLY when
the title and the tree genuinely disagree about what the
object is, or the title alone does not make the object
clear (e.g.  a battery site that could be read as guns,
a bunker, or an installation).  A synonym of your chosen
word is NOT an uncertainty ("dolmen" is not unsure because
"megalithic tomb" exists) -- never list synonyms.  At most
2 entries.  A non-empty list marks the item for human
review, so listing alternatives "to be safe" defeats its
purpose.

confidence -- "high" when the object class is clear from
title + tree, "low" when you filled alternatives or the
title does not make the object clear.  Nothing else.
```

User message:

```
Assign segmentation prompts to these categories:
0. "<title>" | parent: <parent category> | root:
<crawl root> | granularity: <label>
1. ...
```

Root-category screening (Section A). The response for each category gives a yes or no decision and a short reason.

```
You are screening Wikimedia Commons category names to
decide which ones are FAMILIES OF RIGID PHYSICAL OBJECTS
worth crawling for a 3D object-reconstruction dataset.
Each name is a direct subcategory of a broad "umbrella"
category (e.g. "Musical instruments", "Tools", "Furni-
ture", "Containers"); the umbrella is given for context.

Answer YES when the category is a class of manufactured or
carved rigid objects whose members would be photographed
as objects: guitars, pianos, chairs, knives, watches,
telephones, howitzers, loudspeakers, tins, buckets, radars,
electric motors.

Answer NO for anything else:
- topics, activities, concepts, history ("Ball games",
  "History of synthesizers", "Human-computer interaction",
  "Light art", "Synth DIY")
- people and occupations ("Clockmakers", "Sculptors")
- vehicles of any kind -- cars, trucks, ships, aircraft,
  trains, bicycles count as vehicles here ("Vehicles",
  "Anchorhandling, tug and supply ships", "Survey ships")
  -- they are covered by a separate dataset
- soft / deformable / flat things (textiles, rugs, bags,
  clothing, flags, paper, coins, plaques)
- media about objects (photos-of, logos, "Numbers on racing
  vehicles", stickers)
- categories that are groupings by attribute rather than
  a family ("Metal sundials", "Black liftbacks", "...by
  color", "...by year")
- ammunition cartridges are borderline small objects
  -- answer YES for guns/weapons families, NO for
  cartridges/bullets

When the name is a sub-family of something obviously
object-like ("Bullpup", "Guitars by type", "Submachine
guns by model", "Lenses by brand", "Watches by brand")
answer YES -- index suffixes ("by type/brand/model") are
fine; those are how Commons lists the members.

Output for each item: index, decision ("yes"/"no"), and a
reason of at most 6 words.
```

User message:

```
Screen these categories:
0. "<title>" | umbrella: <umbrella category>
1. ...
```

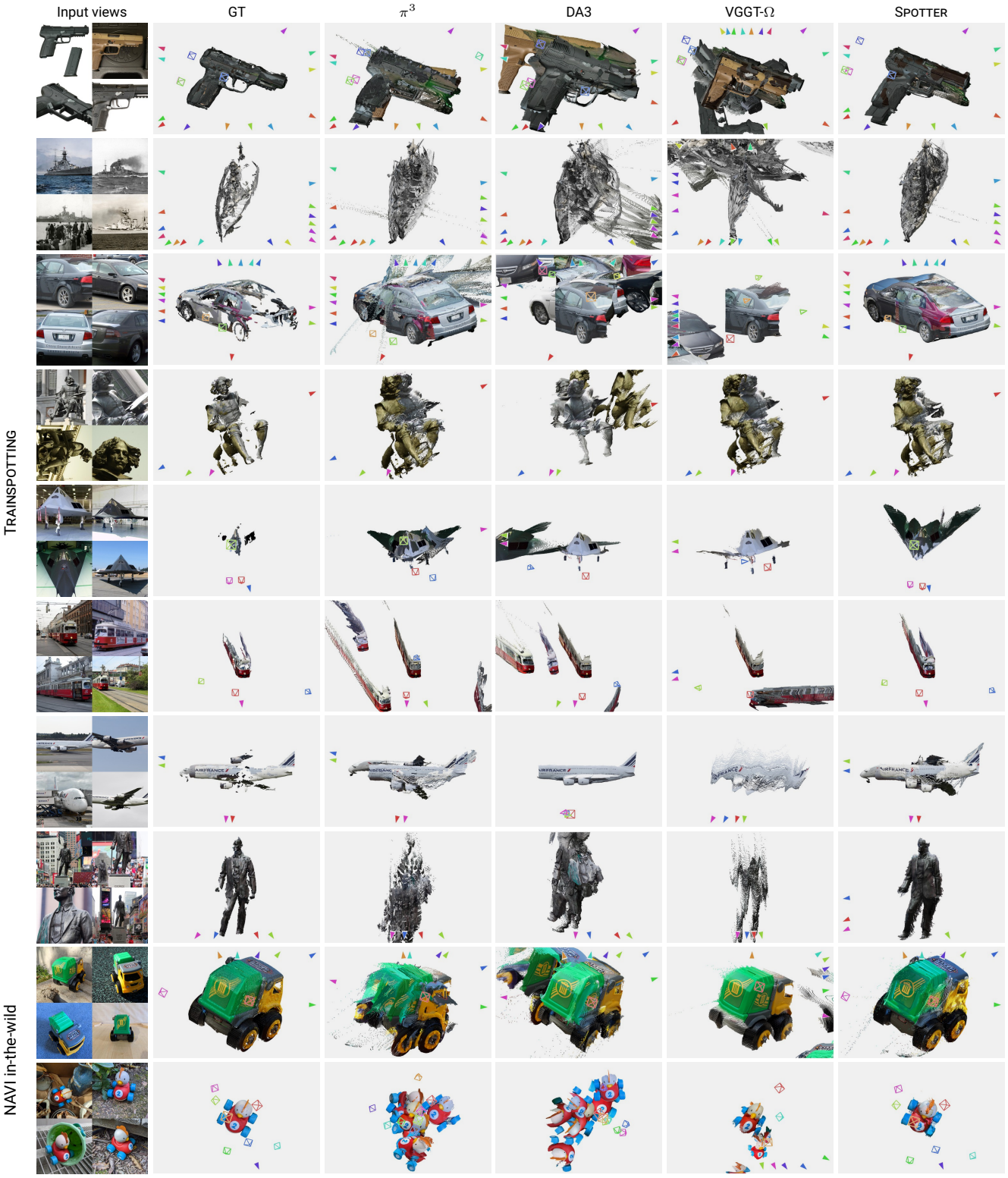


Figure S3. **Additional qualitative comparisons.** Ten held-out samples drawn at random: three TRAINSPOTTING objects with 16 input views, five with 4 input views, and two NAVI in-the-wild objects with 8 input views.

Table S7. **All metrics on TRAINSPOTTING.** Extends Table 3 of the main paper with the median of accuracy and completeness, the inlier ratios at 1.05 and 1.10, and AUC at 15°. Accuracy and completeness are in percent of the object extent.

Method	Point map				Depth				Camera		
	Acc mean	Acc med.	Comp mean	Comp med.	AbsRel	$\delta_{1.03}$	$\delta_{1.05}$	$\delta_{1.10}$	AUC_5	AUC_{15}	AUC_{30}
VGGT [48]	2.38	1.62	8.08	6.48	0.160	32.4	41.8	56.2	11.5	29.9	44.1
π^3 [52]	2.33	1.61	6.22	4.76	0.162	34.7	43.8	56.8	12.5	31.9	46.4
DA3 [25]	2.22	1.60	12.17	10.11	0.189	32.1	41.5	54.6	10.1	21.0	29.4
MapAny. [20]	2.20	1.57	7.75	6.18	0.206	26.6	35.3	49.1	5.3	19.3	33.1
VGGT-Ω [49]	1.84	1.36	12.68	10.77	0.201	28.1	37.1	50.3	7.2	16.9	25.0
SPOTTER	**1.22**	**0.89**	**1.15**	**0.81**	**0.071**	**53.3**	**65.4**	**80.5**	**44.2**	**70.1**	**81.2**

Table S8. **All metrics on the external benchmarks.** The metrics of Table S7 and normal consistency (NC) on the benchmarks of Tables 5 and 6 of the main paper. Accuracy and completeness are in millimeters.

Method	Point map					Depth				Camera		
	Acc mean	Acc med.	Comp mean	Comp med.	NC	AbsRel	$\delta_{1.03}$	$\delta_{1.05}$	$\delta_{1.10}$	AUC_5	AUC_{15}	AUC_{30}
7-Scenes (18 samples)												
VGGT [48]	25.3	11.6	32.6	15.9	0.735	**0.277**	67.5	83.4	92.3	26.6	65.2	81.0
π^3 [52]	**20.0**	10.5	**26.7**	14.3	**0.765**	0.342	70.5	87.1	**96.4**	28.8	68.5	83.3
DA3 [25]	25.4	13.1	35.7	15.9	0.755	0.324	74.1	88.1	94.6	28.5	66.3	81.5
MapAny. [20]	28.2	14.7	36.8	16.0	0.739	0.332	58.3	77.3	93.3	18.2	54.7	72.4
VGGT-Ω [49]	21.8	11.9	27.4	14.4	0.760	0.295	71.9	86.9	94.3	**40.8**	**75.4**	**87.2**
SPOTTER	20.2	**10.2**	27.4	**14.1**	0.760	0.301	**75.1**	**88.3**	94.5	38.8	73.8	86.1
NRGBD (9 samples)												
VGGT [48]	17.4	10.1	16.9	7.0	0.914	0.016	90.9	96.2	98.3	86.4	95.1	97.6
π^3 [52]	17.9	9.3	17.2	6.8	0.913	0.016	88.6	94.6	98.3	86.5	94.7	97.2
DA3 [25]	16.4	7.8	15.3	6.3	**0.932**	**0.010**	**96.7**	**98.0**	98.6	**91.4**	**97.1**	**98.6**
MapAny. [20]	42.8	22.2	30.2	12.5	0.856	0.033	69.2	85.9	95.5	56.5	82.4	90.9
VGGT-Ω [49]	13.3	8.2	9.0	5.7	0.908	0.013	92.4	**98.0**	**99.1**	87.8	95.7	97.9
SPOTTER	**12.9**	**7.7**	**8.3**	**5.5**	0.914	0.013	92.7	97.7	99.0	86.5	95.1	97.3
DTU (22 samples)												
VGGT [48]	1.2	0.7	1.8	1.0	0.699	**0.009**	**97.4**	**98.0**	98.6	**98.1**	**99.4**	**99.7**
π^3 [52]	**1.2**	0.7	**1.5**	**0.6**	0.695	0.014	91.2	96.9	**98.7**	80.0	93.2	96.6
DA3 [25]	1.7	1.0	1.6	0.8	**0.711**	0.009	96.5	97.9	98.6	91.0	97.0	98.5
MapAny. [20]	2.4	1.3	1.9	0.8	0.702	0.067	37.8	55.2	77.8	15.0	53.3	73.4
VGGT-Ω [49]	1.2	**0.7**	1.7	0.6	0.701	0.009	97.1	**98.0**	**98.7**	91.1	97.0	98.5
SPOTTER	1.2	0.7	1.6	0.6	0.700	0.012	95.9	97.5	98.6	86.1	95.3	97.7
ETH3D (13 samples)												
VGGT [48]	291.4	190.7	317.5	204.4	0.841	0.045	52.6	69.5	88.4	37.6	58.5	71.1
π^3 [52]	154.7	87.2	129.7	65.1	0.890	0.032	62.4	78.4	93.4	36.4	58.7	72.9
DA3 [25]	274.3	197.3	270.3	170.1	0.844	0.027	73.2	84.8	95.4	55.4	69.5	75.4
MapAny. [20]	459.7	377.7	542.0	413.9	0.812	0.161	51.9	65.7	81.3	33.8	54.4	65.6
VGGT-Ω [49]	73.6	43.5	61.1	34.5	0.915	**0.017**	**83.5**	**94.5**	**99.2**	**66.9**	**84.9**	**91.8**
SPOTTER	**65.5**	**37.1**	**52.9**	**30.3**	**0.924**	0.017	82.8	94.3	98.7	65.4	84.8	**91.8**
NAVI (281 samples)												
VGGT [48]	2.5	1.6	3.4	2.7	0.751	0.032	71.5	87.1	96.3	53.3	82.7	91.2
π^3 [52]	**2.2**	**1.4**	**3.0**	**2.4**	**0.762**	**0.027**	76.5	91.3	**97.5**	55.8	83.6	91.6
DA3 [25]	3.3	2.0	5.7	4.2	0.682	0.034	**80.0**	**91.9**	97.1	**57.2**	84.1	91.8
MapAny. [20]	3.8	2.3	5.8	4.6	0.663	0.051	56.4	75.3	91.9	28.4	69.1	84.0
VGGT-Ω [49]	2.9	1.8	5.3	4.1	0.692	0.034	76.1	90.5	97.1	55.4	83.6	91.6
SPOTTER	3.1	1.9	4.0	3.0	0.702	0.032	74.5	89.9	97.2	56.7	**84.2**	**91.9**
NAVI (wild) (35 samples)												
VGGT [48]	3.2	2.4	11.5	10.0	0.633	0.085	35.0	48.1	70.8	14.4	34.8	45.5
π^3 [52]	3.7	2.8	9.6	8.1	0.628	0.078	41.2	55.3	77.2	13.1	32.1	43.7
DA3 [25]	3.3	2.5	16.0	14.0	0.578	0.079	41.1	56.9	76.7	12.0	28.5	36.9
MapAny. [20]	4.1	2.9	11.0	9.5	0.588	0.103	29.6	43.4	66.9	3.3	14.7	24.2
VGGT-Ω [49]	**3.1**	2.5	20.4	18.5	0.544	0.101	29.7	44.6	66.9	3.8	11.6	18.8
SPOTTER	3.2	**2.1**	**1.9**	**1.2**	**0.743**	**0.056**	**43.1**	**60.9**	**82.6**	**30.7**	**62.2**	**76.0**