MARTINE: Benchmarking Multi-View 3D Surface Reconstruction Across Viewpoint Coverage, Resolution, and Lighting

Robin Bruneau[1*] **Baptiste Brument**[2*] **Frederic Giraud**[3] **Bastian Sigrist**[3]
Sascha Jecklin[4] **Philipp Fürnstahl**[3] **Bjoern Menze**[1] **Lilian Calvet**[1,3]

[1]University of Zurich, Switzerland [2]IRIT, Université de Toulouse, CNRS, Toulouse, France
[3]Balgrist University Hospital, University of Zurich, Switzerland [4]Independent researcher

Abstract

Controlled acquisition setups are widely used for high-accuracy multi-view 3D surface reconstruction, as they allow sensor configuration, image resolution, viewpoint distribution, and illumination to be specified. Yet benchmarking remains difficult, since reconstruction accuracy depends jointly on the reconstruction paradigm, object properties, and acquisition regime. Existing benchmarks often restrict viewpoint coverage, resolution, or illumination, which can bias comparisons towards methods suited to the chosen setting. Others lack accurate 3D ground truth or quantified image-to-geometry alignment, limiting their use for metric surface evaluation. We introduce MARTINE, a benchmark for controlled 3D object surface reconstruction. MARTINE comprises 28 objects, including transparent materials and fine structures, captured at 9568×6376 pixels from 84 views over a spherical cap under both single-light and multi-light regimes. It provides object-independent camera calibration, accurate 3D ground-truth scans, calibration and image-to-geometry accuracy quantified on test images, and 3D property masks for analysing errors by material properties, visibility, curvature and fine structures. Experiments across photogrammetry, classical multi-view stereo, neural implicit surfaces, Gaussian-splatting-based reconstruction, and multi-view photometric stereo show that MARTINE enables regime-aware evaluation with explicit control over reference-data accuracy. The benchmark resolves sub-millimetric reconstruction differences, with a best observed Chamfer distance of 0.058 mm over the evaluated visible surface.
Project page: `https://martine-benchmark.github.io/`

1 Introduction

3D surface reconstruction is a central problem in computer vision, with applications in cultural heritage preservation, industrial inspection, robotics, augmented reality, and digital content creation. Across these applications, the objective is often not only to recover a visually plausible shape, but to estimate surface geometry with sufficient accuracy for measurement, analysis, or downstream processing. Multi-view RGB surface reconstruction has achieved increasingly accurate results, in particular through advances in photogrammetry software and classical multi-view stereo (MVS) [1–5], neural implicit surface methods [6, 7], and recent multi-view multi-light variants [8, 9]. While 3D reconstruction can be performed in unconstrained environments [10–14], controlled acquisition setups provide important advantages when high accuracy is required. Such setups allow the acquisition conditions to be specified in terms of sensor selection and configuration, image resolution, viewpoint distribution and illumination.

*Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

Table 1: Comparison of object-scale datasets for controlled surface reconstruction. **360° Az.**, **Multi-Elev.**, **Dense Lights**, and **UHD** indicate full azimuthal coverage, multiple elevation levels, dense controlled illumination, and ultra-high-resolution imagery, respectively. **Image–GT Res.** indicates whether object-level residuals between evaluation images and the 3D reference geometry are reported, in addition to camera-calibration or scanner-accuracy statistics. **Challenge Eval.** indicates whether official masks or groups are provided to analyse errors by surface or material challenge. *n.r.* means not reported.

Dataset	360° Az.	Multi-Elev.	Dense Lights	UHD	Image–GT Res.	Challenge Eval.
DTU [15]	×	✓	×	×	*n.r.*	×
LUCES-MV [18]	✓	×	✓	×	*n.r.*	×
DiLiGenT-MV [19]	✓	×	✓	×	*n.r.*	×
Skoltech3D [16]	×	✓	×	✓	*n.r.*	×
OpenSubstance [20]	✓	✓	✓	×	*n.r.*	×
EvalMVX [21]	✓	×	✓	×	*n.r.*	×
Ours	✓	✓	✓	✓	✓	✓

Existing controlled benchmarks cover this acquisition space only partially. DTU [15] and Skoltech3D [16] provide dense but azimuth-limited coverage, rather than broad object-centric coverage for multi-view surface reconstruction. While both include several illumination conditions, the lighting variations are primarily used to evaluate multi-view single-light methods under different static illuminations, rather than to support multi-view photometric stereo, which requires dense illumination sampling. This limitation is quantitatively studied by [17], which evaluates the suitability of existing controlled datasets for photometric-stereo-based surface reconstruction.

Conversely, datasets designed for multi-view photometric stereo combine several viewpoints with multiple controlled illumination conditions, so that surface geometry can be inferred not only from multi-view consistency but also from shading variation. These datasets provide illumination conditions suitable for photometric reconstruction, but each covers the remaining benchmark dimensions only partially, such as viewpoint density, elevation coverage, image resolution, object count, or surface ground truth. DiLiGenT-MV [19] remains the standard benchmark for multi-view photometric stereo, but contains only five objects, each captured from 20 views at 612×512 resolution. More recent datasets increase resolution or object count, but remain limited in viewpoint elevation coverage: LUCES-MV [18] contains 12 views at 2368×1552 resolution, and EvalMVX [21] contains 20 views at 2448×2048 resolution; all three acquire views at a single elevation. Such design choices can also bias cross-paradigm comparisons. EvalMVX [21], for example, compares single-light and multi-light reconstruction using 20 views over 360° azimuth at a single elevation, rather than a denser single-light regime such as the widely used DTU benchmark [15], which provides 49–64 views over approximately 0°–110° azimuth. OpenSubstance [20] substantially improves view and illumination coverage, with 270 viewpoints and more than 1000 lighting conditions, and is a valuable resource for novel-view synthesis, appearance modelling, and object relighting. However, for surface-reconstruction benchmarking, its image-to-geometry alignment relies on object-dependent registration between the ground-truth scan mesh and a COLMAP [22] reconstruction. This can be unreliable when the COLMAP reconstruction is incomplete or geometrically inaccurate (see Figure 1), for example for transparent objects, thin structures, or challenging materials; yet the resulting alignment errors are not systematically quantified or reported. A related concern is addressed by NAVI [14] for in-the-wild object reconstruction, where 3D scanner ground truth is combined with manually verified 2D–3D alignment to obtain accurate object alignments. In addition, OpenSubstance ground-truth scans are not publicly provided for all objects, which limits surface-reconstruction evaluation on challenging cases. Similarly, LUCES-MV [18] does not provide object-to-camera reference alignment for its transparent objects.

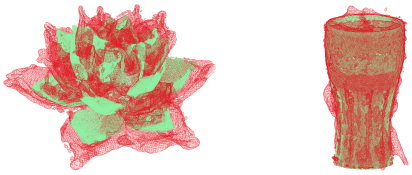


Figure 1: Examples of image–GT misalignment in OpenSubstance [20]. Inaccurate COLMAP reconstructions of an opaque flower (left) and a transparent glass (right) in red. Registered GT scan meshes in green.

A further issue is that the RGB input used by single-light reconstruction methods in multi-view photometric-stereo datasets is often synthesised from the multi-light stack, for example by taking the median [23, 24] or average [18] image over illuminations at each viewpoint. This aggregation has no explicit image-formation basis for single-light reconstruction and is likely motivated by the mitigation of specularities, since specular highlights are often localised and therefore appear only in a subset of the illumination stack. These results remain valuable, but they correspond to a hybrid setting, namely multi-light acquisition followed by single-light reconstruction. They therefore do not directly measure the performance of single-light reconstruction methods under a single-light acquisition regime. As a result, current benchmarks do not jointly provide broad object-centric viewpoint coverage, dense camera-mounted illumination, high-resolution imagery, accurately aligned 3D ground-truth scans, and acquisition regimes that allow single-light and multi-light methods to be compared under conditions designed for each paradigm.

Beyond viewpoint and illumination coverage, image resolution is another source of ambiguity in surface-reconstruction evaluation. This is particularly important for mature photogrammetry pipelines [2, 3], whose patch-based matching procedures can benefit directly from high-resolution imagery. While higher-resolution images provide more information, they do not consistently lead to lower reported reconstruction errors across existing benchmarks. As reported in [17], LUCES-MV [18] and Skoltech3D [16] yield higher errors than DiLiGenT-MV [19], despite substantially higher image resolutions, 2368×1552 and up to 7296×5472, compared to 612×512. This discrepancy is difficult to interpret because resolution is confounded with dataset-specific differences in acquisition, objects, reference scans, and image-to-geometry alignment. It may indicate either that the reference data are not accurate enough for high-resolution evaluation, or that current implementations do not fully exploit high-resolution inputs. Distinguishing between these explanations requires benchmarks with high-resolution imagery, accurate 3D ground truth and quantified image-to-geometry alignment errors.

A further limitation concerns how reconstruction errors are reported. Most benchmarks report aggregate scores per object [15, 16, 18–21], although each object combines several geometric and optical properties. This makes it difficult to isolate failures due to visibility, curvature, fine structures, transparency, specularity, low texture, or other factors. Prior work has already shown that improvements on high-curvature or low-visibility regions can be poorly reflected by global Chamfer-distance or F-score averages [9, 17]. However, such analyses remain scattered across individual papers and are rarely integrated into public benchmark protocols.

To address these limitations, we introduce MARTINE[2], a controlled benchmark for 3D object surface reconstruction. Its main contributions are as follows.

- **A benchmark spanning acquisition regimes.** MARTINE comprises 28 objects placed on a motorised turntable and imaged with a high-end camera mounted on a robotic arm. For each object, we acquire 84 views over a spherical cap at 9568×6376 resolution under two lighting regimes. The single-light regime uses a reflecting cone and a softbox flash to approximate rotation-invariant ambient illumination around the turntable axis, while the multi-light regime uses 30 camera-mounted LED light sources activated sequentially. This design provides acquisition conditions suited to dense-view single-light reconstruction as well as multi-view multi-light reconstruction. The dataset includes challenging geometries and materials, including fine structures, transparent objects, low-texture surfaces, and specular or glossy regions.
- **Reference data with quantified accuracy.** MARTINE provides object-independent camera calibration, calibration accuracy measured on test views, accurate 3D ground-truth scans, and easy-to-interpret 2D–3D registration residuals on test images. These elements make it possible to evaluate surface reconstruction while explicitly accounting for reference-data accuracy.
- **Challenge-level performance analysis.** We provide evaluation code and 3D ground-truth masks to report reconstruction errors by optical properties, visibility, curvature and fine structures, rather than only by object identity.

[2]Dataset: `https://doi.org/10.7910/DVN/2BEPEH`; download and conversion tools: `https://github.com/martine-benchmark/martine-dataset`.

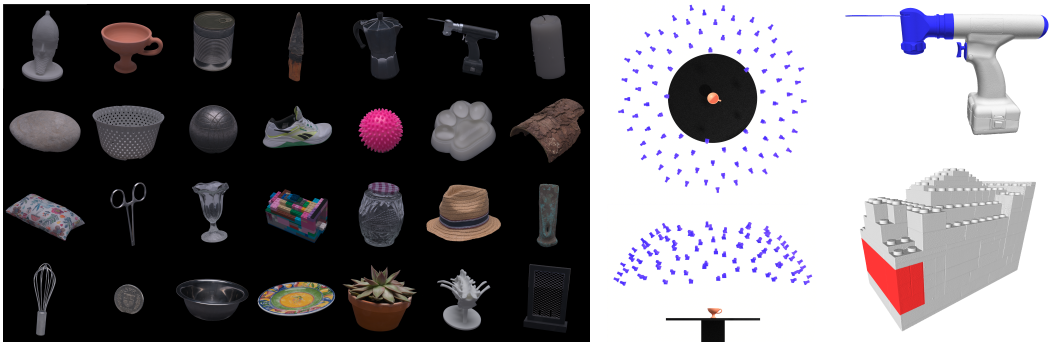
- **A first cross-paradigm evaluation.** We provide an initial evaluation of representative methods from photogrammetry, patch-based MVS, 3D Gaussian Splatting-based reconstruction, neural implicit surfaces, and multi-view photometric stereo.

Together, these contributions make MARTINE a benchmark for studying how acquisition regime, object properties, and reconstruction paradigm jointly affect surface-reconstruction accuracy.

2 Dataset and Acquisition Design

2.1 Overview

Objects. The object set is designed to cover a range of geometric and optical properties relevant to image-based surface reconstruction. It includes simple textured and matte objects, as well as more challenging cases with fine structures, deep concavities, transparent objects, low-texture surfaces, and highly specular materials. Objects are shown in Figure 2(a).



(a) (b) (c)

Figure 2: Dataset overview. (a) All objects in the proposed dataset, showing high geometric and optical diversity, including low-texture regions, fine structures, transparent and specular materials, and high-frequency geometric details. (b) Dense viewpoint sampling with wide angular coverage, obtained by coupling a robotic arm with a turntable. (c) Two examples of ground-truth meshes with manually curated challenge-level 3D masks highlighting specular regions (top, in blue) and transparent regions (bottom, in red) for challenge-level evaluation.

Viewpoints. The proposed viewpoints follow a Fibonacci-sphere distribution restricted to a spherical cap of 80 cm radius around the object. We acquire 84 views between 20° and 70° elevation, providing near-uniform sampling of the cap. Each view is captured at 9568×6376 resolution with a Sony α7R V camera. Viewpoints are shown in Figure 2(b).

Single-light setup. The single-light regime uses an ambient illumination setup designed to be approximately rotationally symmetric around the turntable axis. This yields nearly view-invariant lighting for the object volume above the turntable, and provides favourable conditions for standard multi-view reconstruction methods. In particular, it substantially reduces the view-dependent appearance changes that would occur if the object were rotated under a fixed asymmetric light source. See Figure 3(a). The rotational symmetry of the illumination was evaluated quantitatively by analysing intensity variations along the isophotes of a white 25 cm sphere placed at the centre of the turntable. The measured standard deviation across viewpoints ranges from 0.5% to 2.7% of the intensity range (mean 1.4%). See details in the appendix (Section C).

Multi-light setup. The multi-light regime uses a rig of 30 LEDs rigidly mounted around the camera. This number of illumination conditions is consistent with the quantitative evaluation of the state-of-the-art photometric stereo method UniMS-PS [25], which reports no improvement from increasing the number of input lights beyond 30 in their setting. It also reflects a practical trade-off in our acquisition design, as each additional light source multiplies the number of images, while the dataset already provides high-resolution 16-bit imagery, dense viewpoints, and two acquisition regimes. This design provides dense illumination for multi-view multi-light reconstruction. See Figure 3(b).

Challenge-level masks. We provide 3D masks on the ground-truth geometry to support evaluation beyond object-level averages. The masks identify Lambertian or matte regions, textureless regions, transparent parts, specular regions, high-curvature regions and fine structures. The material and structure masks (Lambertian, textureless, transparent, specular, fine structure) are segmented manually on the ground-truth mesh, as conservatively as possible: only regions whose property is unambiguous are labelled, and regions where the label was uncertain are excluded from the evaluation. High curvature is computed automatically from the ground-truth geometry. We also provide visibility levels, defined from the number of views observing each surface point. Low-visibility regions generally correspond to concavities or self-occluded areas, and can also be affected by inter-reflections. Examples of 3D masks are shown in Figure 2(c).

2.2 Acquisition Setup

View acquisition. The setup incorporates a Kuka LBR Med 14 industrial robotic arm and a Noxon 360° turntable with an 80 cm diameter. The combination of these two tools allows us to define viewpoints derived from a Fibonacci sphere. This specific trajectory is achieved physically through the ROS 2-managed synchronisation of 10 turntable rotations and precise robotic arm movements (see Figure 3). The turntable is reset between the single-light and multi-light acquisition sequences. This step is intended to maintain consistency between the two capture modes.

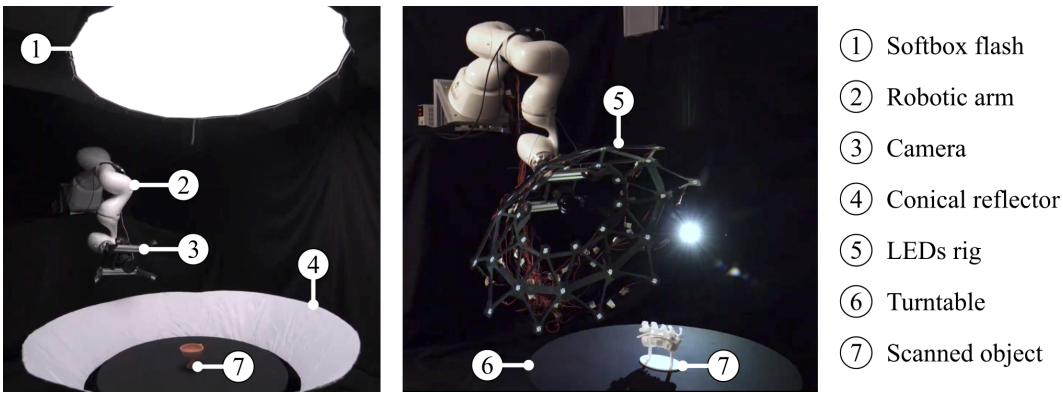


Figure 3: Acquisition setup. (a) Single-light acquisition regime, using an ambient illumination setup designed to be rotation-invariant around the turntable axis. (b) Multi-light acquisition regime, using a rig of 30 LEDs rigidly mounted around the camera. At each viewpoint, the LEDs are activated sequentially while the camera pose remains fixed.

Lighting setup. We acquire each object under two illumination regimes (Figure 3(a) and (b)). In the single-light regime, the setup uses a 120 cm softbox flash together with a reflecting cone placed around the turntable to produce an approximately rotation-invariant ambient illumination around the object. In the multi-light regime, illumination is provided by a spherical-cap rig containing 30 Luxeon Rebel LEDs of 220 lm each. The 85 cm-diameter rig is rigidly mounted to the camera, ensuring a fixed relative configuration between the camera and the light sources across viewpoints. The LEDs cover a range of lighting directions around the optical axis while remaining at approximately the same distance from the object. Acquisition is performed in a darkroom, and the LEDs are activated sequentially following a one-light-at-a-time protocol. A black image per viewpoint is also captured for the calibration.

2.3 Ground Truth Acquisition

We acquire reference geometry with an Artec Spider optical scanner [26], specified at 0.05 mm 3D point accuracy and 0.1 mm 3D resolution. As in high-accuracy object-reconstruction benchmarks such as Skoltech3D [16], specular and transparent objects are sprayed with Aesub Blue, a sublimating opacifying spray, to reduce optical artefacts during scanning.

The scans are acquired manually rather than with a fixed scanner trajectory. This choice is more time-consuming, but gives the operator direct control over the sensor path and allows coverage to be checked continuously during acquisition. It is particularly useful for objects with concavities,

thin structures, or self-occluded regions, where a fixed scanner and turntable setup may leave parts of the surface insufficiently observed. In total, 21 of the 26 ground-truth meshes are obtained with the Artec Spider scanner. For objects with narrow concavities or thin structures that are difficult to observe with optical scanning (PLANT, GRID, and SPINE), or for which the opacifying spray proved ineffective (PAW PAD and BOWL), the ground truth is instead obtained from computed tomography scans acquired with a NAEOTOM Alpha CT scanner at a voxel size of 0.117 mm; for the WARRIOR HEAD, the Artec and CT scans are fused. The scans are then segmented using Materialise Mimics 26.0 [27].

This workflow follows a monitored pipeline in which each reference model is inspected manually. Among the 28 objects, two do not have ground-truth scans. The first is PILLOW, because its surface texture is partly interpreted by the Artec Spider as geometric variation. The second is COIN, because neither scanning modality could capture its fine geometric details reliably. Both objects are nevertheless released with their images, object masks and calibrated cameras, but they are excluded from the evaluation: all quantitative results are reported on the remaining 26 objects.

2.4 Camera Calibration and Scan Alignment

This section describes the reference-data pipeline used to place all images and ground-truth scans in a common metric coordinate system. We first present the camera calibration and its validation, then describe the scan-alignment procedure and its validation.

Camera calibration. The camera-calibration procedure is designed to be object-agnostic while providing quantified accuracy. Object-agnostic calibration is important because camera poses should not depend on the geometry, texture, or material of the reconstructed object. Otherwise, pose errors could vary with object difficulty, making reconstruction errors harder to interpret.

Several strategies are possible. One option is to rely on the robot and turntable poses, but prior work [28] has shown that this can be insufficiently accurate at this image resolution. Another option, adopted by EvalMVX [21], is to place a calibration board below the object. This provides image features for pose estimation, but is not well suited to calibrated photometric stereo because the checkerboard can introduce inter-reflections and make object-contour extraction less robust during the subsequent alignment procedure. A related strategy, used in OpenSubstance [20], places markers on a vertical cylindrical support below the object, which likely reduces inter-reflections compared with a planar checkerboard. In contrast, our calibration uses only the textured black fabric covering the turntable, with the object and background excluded from feature extraction. In our evaluation, this provides the lowest and most consistent reprojection errors on a common set of test corners, while avoiding inter-reflections from markers.

Concretely, all images from each acquisition are calibrated jointly in a single Structure-from-Motion (SfM) run. This includes the single-light images, the multi-light images acquired with all LEDs active, and the opacifying-spray images when available. We use Segment Anything Model segmentation to generate turntable masks that restrict feature extraction to the black fabric region. SfM is performed with Meshroom [3], which builds on OpenMVG [29], using DSP-SIFT [30] features. Additional details on the calibration are provided in the appendix (Section A).

Calibration accuracy. We compare the proposed calibration procedure with three alternatives, some of which are representative of existing benchmarks. The first is COLMAP [4] applied to the same textured cloth. The second uses a ChArUco board [31], analogous to the ArUco-board calibration used with Metashape in EvalMVX [21]. The third uses an ArUco cylinder placed on the turntable base, analogous to the ARTag fiducial markers [32] used in OpenSubstance [20]. We use ArUco markers [31] instead of ARTag markers, as they have been reported to provide better performance in prior evaluations [33, 34]. Accuracy is evaluated on an 18×18-corner checkerboard placed on the turntable, which is visible in the calibration images but excluded from feature extraction by the segmentation masks, so that its corners are never used for calibration. We measure the reprojection error of the checkerboard corners on the 20 views with the largest viewing angle with respect to the board plane ($55.6°$–$89.0°$), the same selection for all four configurations, so that all of them are evaluated on the same 6 480 corners. Our SfM-based approach achieves the lowest reprojection error, with a mean of 0.542 px (median 0.524 px), corresponding to approximately 0.04 mm at the nominal object distance. It is followed by the ArUco cylinder with a mean of 0.666 px (median 0.542 px), COLMAP with 0.809 px (0.799 px), and ChArUco with 1.086 px (0.924 px). Note that

the measured residual includes checkerboard-corner localisation error in addition to calibration error. Full distributions are reported in the appendix (Table 3).

Scan alignment. We formulate scan alignment as a contour-based ICP problem in the image domain.[3] The goal is to align the ground-truth mesh with the calibrated image sequence by matching projected mesh silhouettes to object contours extracted in the images. Since the ground-truth scans are expressed in metric units, while SfM camera poses are recovered up to an unknown scale, the alignment estimates a 7-DoF similarity transform consisting of rotation, translation, and uniform scale. The inverse transform is then applied to the camera poses, expressing them in the metric coordinate frame of the ground-truth scan and yielding a metrically consistent image-to-geometry alignment. Target contours are extracted from single-light images using Canny edge detection within a local search band defined from SAM object masks, with contour-extraction parameters optimised per object. For specular or transparent objects, we use images captured after applying the white sublimating opacifying spray, which increases the contrast between the object and the black turntable near the silhouette. The registration then optimises the similarity transform by minimising the Euclidean distance from the target image contour points to the projected silhouette of the ground-truth mesh. The mathematical formulation of the optimisation and the selection of the contour-extraction parameters are provided in the appendix (Section B).

Alignment accuracy. The alignment is validated on control views excluded from registration. For each object, we select eight control views by k-means clustering of the camera centres. At each iteration, the similarity transform is estimated from the remaining views using the object-specific contour-extraction parameters, and the reprojection residual is measured on the excluded view. We report the average residual over the eight leave-one-control-view evaluations, yielding a pixel-level image-to-geometry error at the original sensor resolution. The mean LOO residual across all objects is 1.22 px, with a median of 1.20 px, and all 26 objects achieve a mean residual below 2 px, as shown in Table 2. We report reprojection residuals rather than silhouette IoU, as used in EvalMVX [21], because preliminary tests showed that IoU on downsampled masks can be insensitive to small contour misalignments. Additional details are provided in the appendix (Figure 9).

Table 2: Leave-one-out registration residuals (mean reprojection error in pixels, over 8 control views per object.

ROCK	VASE	HAMMER HEAD	CANDLE	CAN	GLASS	PETANQUE BALL	JAM POT
1.06	1.08	1.02	0.78	1.21	1.14	0.84	1.26

COLANDER	WOOD	PLATE	PAW PAD	MOKA POT	LEGO	SCISSORS	KNIFE	SPIKES
1.11	1.45	1.24	1.01	1.98	1.35	0.89	1.10	1.11

STRAW HAT	SNEAKERS	SAW	WHISK	WARRIOR HEAD	PLANT	GRID	BOWL	SPINE
1.36	1.27	1.09	1.18	1.27	1.31	1.28	1.69	1.63

Taxonomy of reconstruction challenges. Existing benchmarks typically report one aggregate score per object, even though a single object may contain several distinct surface properties, such as matte regions, low texture, high curvature, transparency, or specularity. As a result, an object-level score does not indicate which properties drive the reconstruction error. To make this analysis possible, we annotate each ground-truth mesh with per-vertex labels describing relevant geometric and optical properties. The seven property labels are defined as follows. *Lambertian* denotes matte regions whose appearance is approximately diffuse. *Specular* denotes regions with strong view-dependent reflectance, including glossy or metallic surfaces. *Transparent* denotes transmissive regions, for which image evidence can be affected by refraction, reflection, or background visibility. *Textureless* denotes regions with absent surface texture, which provide very limited local image correspondences. *Fine structure* denotes thin geometric elements with small cross-section, such as wires, blades and spikes. *High curvature* denotes vertices with large local maximum principal curvature ($|k| > 0.75$ mm^{-1}), estimated using the Voronoi Covariance Measure ($R = 0.5$ mm, $r = 1$ mm) on a densified ground-truth mesh (edge splitting > 0.15 mm). *Low visibility* denotes vertices observed in only a small number of calibrated views, thresholded at 1 to 3 views in our experiments. This property-level evaluation is particularly useful for regions that occupy only a small fraction of the object surface. Errors on such under-represented regions can be diluted in global metrics, such as mean F-score or Chamfer distance over the full object.

[3]Code: `https://github.com/martine-benchmark/contour-registration`.

3 Experiments

3.1 Baselines

We evaluate representative methods from the reconstruction paradigms supported by the dataset. For photogrammetry, we use Meshroom [3], an open-source pipeline based on CMPMVS [35], and RealityScan [2], a widely used closed-source software package. As a classical MVS baseline, we use COLMAP [4, 22]. We also evaluate PGSR [36] for 3D Gaussian Splatting-based surface extraction, and two neural implicit surface reconstruction methods, NeuS2 [37] and Neuralangelo [7]. For multi-view photometric stereo, we evaluate SuperNormal [8] and RNb-NeuS2 [17], using the state-of-the-art photometric stereo method UniMS-PS [25] to estimate per-view normal maps from the multi-light images. Method settings are given in the appendix (Section E.4). For the per-object and per-property analyses, each family is represented by one method: RealityScan, COLMAP, PGSR, NeuS2 and RNb-NeuS2. For neural implicit surface reconstruction, we decide to use NeuS2 as the representative method because Neuralangelo takes about 160 times as long to run at quarter resolution.

3.2 Evaluation Protocol

Image resolutions. Each method is evaluated at up to four image resolution levels: full resolution (9568×6376), half resolution (4784×3188), quarter resolution (2392×1594), and one-eighth resolution (1196×797). Only the two photogrammetry pipelines (Meshroom and RealityScan) process full-resolution images: all other methods fail at this resolution for lack of GPU memory and are evaluated at lower resolutions. RealityScan is run at all four levels; COLMAP, PGSR, NeuS2 and RNb-NeuS2 are run at half, quarter and one-eighth resolution, and Neuralangelo and SuperNormal at quarter resolution only. View subsets of 40, 20, 10 and 5 views, selected by k-means, are evaluated at quarter resolution for RealityScan, COLMAP, PGSR, NeuS2 and RNb-NeuS2 (Table 7 in the appendix).

Evaluation protocol. Before evaluation, reconstructed meshes are filtered by removing vertices not visible from any camera, and unreliable regions of the ground-truth meshes (e.g., mounting bases) are excluded using per-object 3D masks; both are then cleaned and resampled identically. Surface accuracy is measured using precision, recall, and F-score curves for distance thresholds between 0 and 4 mm.[4] Precision (accuracy) is the fraction of reconstructed points within a given distance of the reference surface, recall (completeness) the fraction of reference points within that distance of the reconstruction, and the F-score their harmonic mean. Full details are given in Section E.1 of the appendix.

3.3 Results

Figure 4 reports precision, recall, and F-score curves pooled over all objects at quarter resolution. In this configuration, the neural implicit methods reach the highest F-scores (77.9% for Neuralangelo and 75.5% for NeuS2 at 0.5 mm), followed by COLMAP (70.6%), whose precision is the highest beyond 0.6 mm but whose recall is lower. The multi-light methods do not outperform their single-light counterparts: RNb-NeuS2 (66.7%) has a higher recall than COLMAP but a lower precision, and remains below NeuS2, on which it builds, while SuperNormal ranks last (60.8%). For the present benchmark and implementations, the photometric cues estimated from the multi-light images therefore do not translate into a higher global accuracy, which indicates that further work is needed to fully exploit photometric-stereo information in multi-view surface reconstruction.

Figure 5 studies the effect of reducing the number of views and the input resolution on the VASE object; it illustrates the per-object analyses enabled by the benchmark rather than supporting general conclusions. On this object, COLMAP and PGSR degrade under viewpoint sparsity and fail with 5 views, whereas NeuS2 and RNb-NeuS2 remain more stable, RNb-NeuS2 performing best in the sparse-view setting, possibly thanks to the normal cues provided by photometric stereo. Reducing the image resolution mostly affects the photogrammetry and MVS pipelines: RealityScan drops from 99.3% at full resolution to 28.0% at one-eighth resolution, and COLMAP fails at one-eighth resolution. PGSR also fails at one-eighth resolution, whereas NeuS2 and RNb-NeuS2 are almost

[4] Evaluation code: `https://github.com/martine-benchmark/eval-3d`.

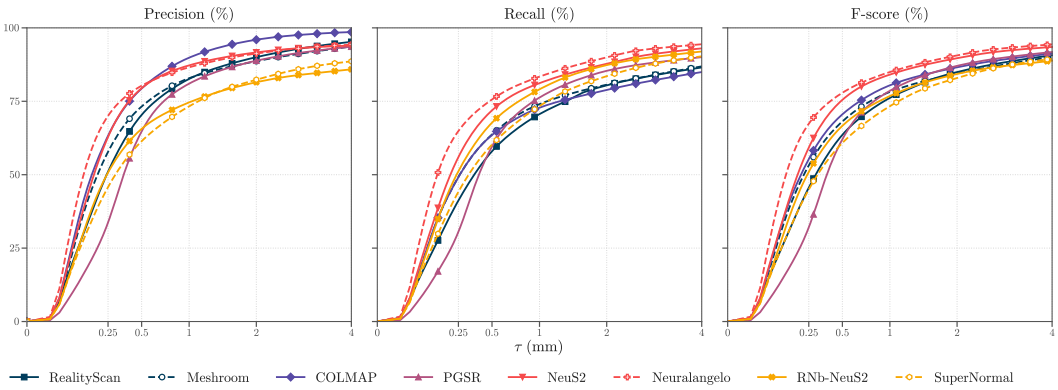


Figure 4: Global precision, recall, and F-score curves pooled across all evaluated objects at quarter resolution (d4). Each curve aggregates all point-to-surface distances before computing the metric, so larger objects contribute proportionally more.

unaffected. Across the 26 objects (Table 7 in the appendix), COLMAP and PGSR reconstruct 14 and 13 objects with 10 views and 4 and 3 with 5 views, while NeuS2 and RNb-NeuS2 reconstruct all of them in every configuration.

Additional exemplar results are shown in Figure 6, which provides spatial error maps for representative objects and methods.

The property-level radar plot in Figure 7 compares the representative methods on the same ground-truth surface samples for each property. Transparent and low-visibility regions yield the lowest F-scores for all methods except Neuralangelo, which scores lowest on low-visibility regions and fine structures (38.2%). NeuS2 leads on specular and transparent regions, RNb-NeuS2 on high-curvature and low-visibility regions, and Neuralangelo on Lambertian and textureless ones, narrowly ahead of COLMAP on the latter (65.3% vs. 63.5%). NeuS2 and RNb-NeuS2 are on par on fine structures (92.1% and 91.0%), well above COLMAP (59.6%) and RealityScan (52.3%). These observations should be read with caution: several gaps are small, and the analysis is conducted at a single resolution. Patch-based MVS such as COLMAP relies on local photo-consistency across views, so it is expected to be strongest where this evidence is stable and to struggle where the assumption is challenged, as on specular or transparent surfaces. Neural surface methods instead optimise a global, regularised volumetric representation, which better tolerates the view-dependent appearance of such regions, and their pixel-wise optimisation has been shown to better preserve thin structures and high-frequency detail [6, 38]. Confirming these trends requires testing at different resolutions and view counts. The proposed benchmark is designed for such comparisons.

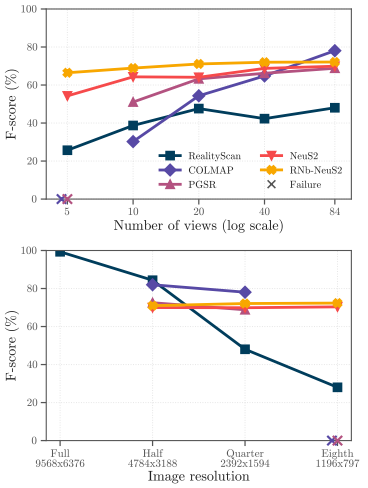


Figure 5: F-score at 1 mm on the VASE object versus number of input views on a log scale (top) and image resolution (bottom). Crosses plotted at 0% F-score mark failed reconstructions in the 5-view and one-eighth-resolution settings.

4 Conclusion and Future Work

In this work, we introduced MARTINE, a benchmark for object surface reconstruction in controlled environments designed to address limitations of existing datasets. MARTINE provides 28 objects, including challenging cases such as transparent materials and fine geometric structures, and controls acquisition parameters such as viewpoint distribution, image resolution, and illumination. This enables reconstruction paradigms to be compared on the same objects under well-defined acquisition regimes. The emphasis on reference-data accuracy, through object-independent calibration and quantified image-to-geometry residuals, supports a clearer interpretation of reconstruction errors. In

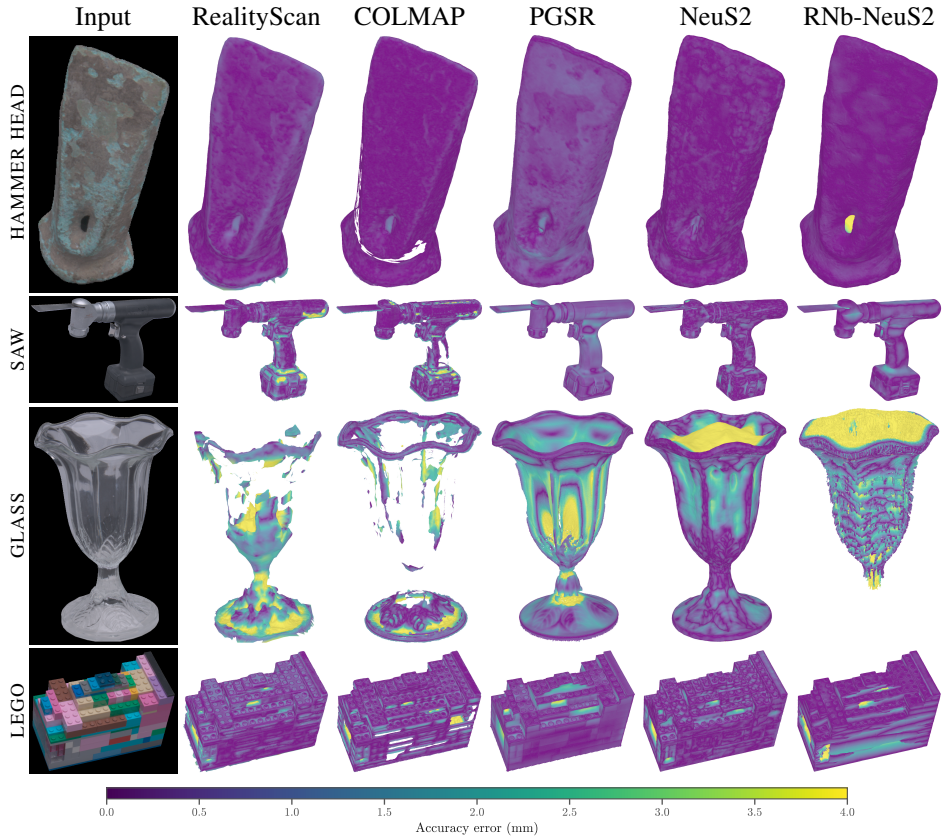


Figure 6: Qualitative comparison of accuracy error heatmaps (viridis colour map, 0–4 mm, larger errors saturated) at quarter resolution (d4).

addition, 3D property masks enable a more detailed analysis of errors across surface regions and object properties.

The current version of the dataset nevertheless has limitations. First, our property taxonomy is partly qualitative, and the manual masks label only unambiguous regions. Future work could use objective criteria to automate annotation and cover a wider range of surfaces. Where a property can be measured on a continuous scale, evaluation could show how reconstruction performance varies across that scale. Second, this work focuses on surface reconstruction. However, the dataset is directly usable for novel-view synthesis, relighting, and inverse rendering, even though we do not evaluate these tasks. Third, the time required to acquire and validate the data limits the benchmark to 28 objects. Each object is captured in 2 604 images at 61 megapixels, and the full workflow takes about one day per object, including image acquisition, scanning, alignment, and quality control. We adopted this protocol to prioritise the accuracy of the reference data.

More generally, we hope that this benchmark will encourage the community to characterise reconstruction methods by the technical challenges they address, rather than only through aggregate rankings. We also hope it will promote more systematic reporting of ground-truth accuracy, camera-calibration accuracy, and image-to-geometry alignment errors for this type of benchmark.

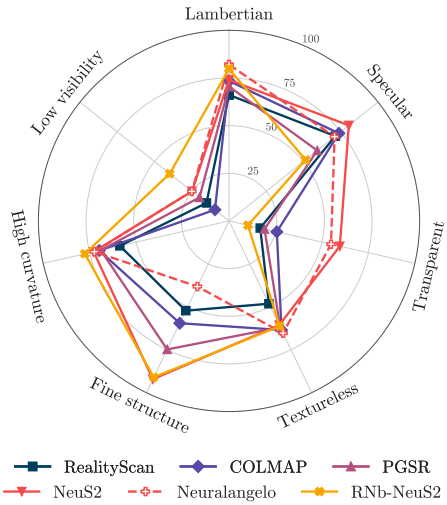


Figure 7: Radar chart of F-score at 0.5 mm per reconstruction challenge, at quarter resolution, for the five representative methods and Neuralangelo. Higher is better.

Acknowledgments and Disclosure of Funding

Robin Bruneau was funded by the SNSF SINERGIA (F01-10266) and the Helmut Horten Foundation. Baptiste Brument was funded by the OPEN-DOPAMIn project (CNRS Innovation). This work has been supported by the OR-X—a Swiss national research infrastructure for translational surgery—and associated funding by the University of Zurich and University Hospital Balgrist. This work was granted access to the Juliet computing platform, supported by a French government grant managed by the Agence Nationale de la Recherche under the "Investissements d'avenir" programme (reference "ANR-21-ESRE-0051"). The authors declare no competing interests.

References

[1] Agisoft LLC. Agisoft Metashape Professional. Photogrammetry software, 2026. URL `https://www.agisoft.com/`.

[2] Capturing Reality / Epic Games. RealityScan. Photogrammetry software (formerly RealityCapture). Version used: 2.1.0, 2025. URL `https://www.realityscan.com/`.

[3] C. Griwodz, S. Gasparini, L. Calvet, P. Gurdjos, F. Castan, B. Maujean, et al. AliceVision Meshroom: An Open-Source 3D Reconstruction Pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pages 241–247. ACM, 2021.

[4] J. L. Schönberger and J.-M. Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[5] J. Wu, R. Li, H. Xu, W. Zhao, Y. Zhu, J. Sun, et al. GoMVS: Geometrically Consistent Cost Aggregation for Multi-View Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20207–20216, 2024.

[6] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[7] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, et al. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[8] X. Cao and T. Taketomi. SuperNormal: Neural Surface Reconstruction via Multi-View Normal Integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20581–20590, 2024.

[9] B. Brument, R. Bruneau, Y. Quéau, J. Mélou, F. Lauze, J.-D. Durou, et al. RNb-NeuS: Reflectance and Normal-Based Multi-View 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[10] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, et al. A Multi-View Stereo Benchmark with High-Resolution Images and Multi-Camera Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[11] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

[12] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, et al. BlendedMVS: A Large-scale Dataset for Generalized Multi-view Stereo Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.

[13] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny. Common Objects in 3D: Large-scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.

[14] V. Jampani, K.-K. Maninis, A. Engelhardt, A. Karpur, K. Truong, K. Sargent, et al. NAVI: Category-Agnostic Image Collections with High-Quality 3D Shape and Pose Annotations. In *Advances in Neural Information Processing Systems*, 2023.

[15] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanæs. Large Scale Multi-view Stereopsis Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014.

[16] O. Voynov, G. Bobrovskikh, P. Karpyshev, S. Galochkin, A.-T. Ardelean, A. Bozhenko, et al. Multi-Sensor Large-Scale Dataset for Multi-View 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21392–21403, 2023.

[17] R. Bruneau, B. Brument, Y. Quéau, J. Mélou, F. B. Lauze, J.-D. Durou, et al. Multi-View Surface Reconstruction Using Normal and Reflectance Cues. *International Journal of Computer Vision*, 134:69, 2026.

[18] F. Logothetis, I. Budvytis, S. Liwicki, and R. Cipolla. LUCES-MV: A Multi-View Dataset for Near-Field Point Light Source Photometric Stereo. arXiv preprint arXiv:2412.16737, 2024.

[19] M. Li, Z. Zhou, Z. Wu, B. Shi, C. Diao, and P. Tan. Multi-View Photometric Stereo: A Robust Solution and Benchmark Dataset for Spatially Varying Isotropic Materials. *IEEE Transactions on Image Processing*, 29(1):4159–4173, 2020.

[20] F. Pei, J. Bai, X. Feng, Z. Bi, K. Zhou, and H. Wu. OpenSubstance: A High-quality Measured Dataset of Multi-View and -Lighting Images and Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5221–5231, 2025.

[21] Z. Yang, J. Ren, X. Wang, Z. Li, X. Cao, H. Guo, et al. EvalMVX: A Unified Benchmarking for Neural 3D Reconstruction under Diverse Multiview Setups. arXiv preprint arXiv:2602.24065, 2026.

[22] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Proceedings of the European Conference on Computer Vision*, pages 501–518, 2016.

[23] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool. Uncertainty-Aware Deep Multi-View Photometric Stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12601–12611, 2022.

[24] B. Kaya, S. Kumar, C. Oliveira, V. Ferrari, and L. Van Gool. Multi-View Photometric Stereo Revisited. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3126–3135, 2023.

[25] C. Hardy, Y. Quéau, and D. Tschumperlé. Uni MS-PS: A Multi-Scale Encoder-Decoder Transformer for Universal Photometric Stereo. *Computer Vision and Image Understanding*, 248:104093, 2024.

[26] Artec 3D. Artec 3D Scanner. `https://www.artec3d.com/portable-3d-scanners/`.

[27] Materialise. Materialise Mimics. `https://www.materialise.com/en/healthcare/mimics`.

[28] E. Most, J. Hein, F. Giraud, N. A. Cavalcanti, L. Zingg, B. Brument, et al. Acquiring submillimeter-accurate multi-task vision datasets for computer-assisted orthopedic surgery. *International Journal of Computer Assisted Radiology and Surgery*, 20(6):1293–1300, 2025.

[29] P. Moulon, P. Monasse, R. Perrot, and R. Marlet. OpenMVG: Open Multiple View Geometry. In *Proceedings of the International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

[30] J. Dong and S. Soatto. Domain-Size Pooling in Local Descriptors: DSP-SIFT. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5097–5106, 2015.

[31] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez. Automatic Generation and Detection of Highly Reliable Fiducial Markers under Occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[32] M. Fiala. ARTag, a Fiducial Marker System Using Digital Techniques. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 590–596, 2005.

[33] G.-C. Patru, D. Rosner, and R.-V. Rughinis. Fiducial Marker Systems Overview and Empirical Analysis of ArUco, AprilTag and CCTag. *UPB Scientific Bulletin, Series C*, 2023.

[34] A. Zakiev, K. Shabalina, T. Tsoy, and E. Magid. Pilot Virtual Experiments on ArUco and ARTag Systems Comparison for Fiducial Marker Rotation Resistance. In *Proceedings of the 14th International Conference on Electromechanics and Robotics "Zavalishin's Readings"*, volume 154 of *Smart Innovation, Systems and Technologies*, pages 455–464. Springer, 2020.

[35] M. Jancosek and T. Pajdla. Multi-View Reconstruction Preserving Weakly-Supported Surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3121–3128, 2011.

[36] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, et al. PGSR: Planar-based Gaussian Splatting for Efficient and High-Fidelity Surface Reconstruction. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[37] Y. Wang, Q. Han, M. Habermann, K. Daniilidis, C. Theobalt, and L. Liu. NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[38] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman. Volume Rendering of Neural Implicit Surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[39] Calib.io. Checkerboard Targets. `https://calib.io/products/charuco-targets?variant=9400454905903`.

[40] Q. Mérigot, M. Ovsjanikov, and L. J. Guibas. Voronoi-Based Curvature and Feature Estimation from Point Clouds. *IEEE Transactions on Visualization and Computer Graphics*, 17(6):743 – 756, 2011.

A Camera Calibration Specifications

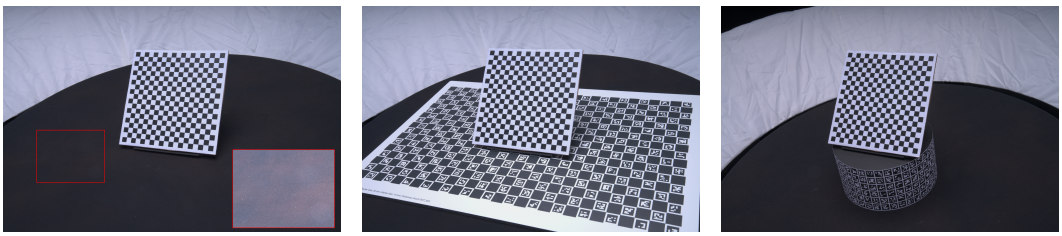
A.1 Calibration Accuracy

We compare our SfM-based camera calibration approach against three alternatives representative of existing benchmarks: COLMAP, a ChArUco board (similar to EvalMVX [21]), and an ArUco cylinder (similar to OpenSubstance [20]). As discussed in the main paper, accuracy is measured by the reprojection error on a test checkerboard that appears in the calibration images but is excluded from feature extraction by the segmentation masks. Figure 8 illustrates these different calibration scenarios.

Table 3 reports the distribution of the per-corner residuals using the 20 views with the largest viewing angle with respect to the board plane ($55.6°$–$89.0°$). Across all four configurations, the residuals are computed over the same 6 480 corners. Our method obtains the lowest mean (0.542 px, about 0.04 mm) and the shortest tail (95th percentile 0.931 px, maximum 1.489 px), whereas its median is close to that of the ArUco cylinder (0.524 against 0.542 px).

Table 3: Distribution of the per-corner reprojection residual on the test checkerboard on the 20 frontal views, in pixels at the full 9568×6376 resolution.

Calibration	n	Mean	Median	RMS	IQR	P95	Max
Textured cloth (ours)	6 480	0.542	0.524	0.584	0.298	0.931	1.489
ArUco cylinder [20]	6 480	0.666	0.542	0.824	0.555	1.621	4.010
COLMAP	6 480	0.809	0.799	0.913	0.612	1.559	2.917
ChArUco board [21]	6 480	1.086	0.924	1.312	0.865	2.541	5.199



(a) Textured cloth (ours) (b) ChArUco board (c) ArUco cylinder

Figure 8: Calibration setup comparison. All configurations include a test checkerboard used only for evaluation. (a) Our approach uses SfM on the textured black cloth. The inset shows a brightness-enhanced crop of the cloth micro-texture, which provides point correspondences at the ultra-high image resolution. (b) ChArUco board [39]. (c) ArUco cylinder on the turntable base.

A.2 SfM Configuration Details

Camera calibration is performed with Meshroom [3] using DSP-SIFT [30] features, which improve robustness to scale changes compared to standard SIFT by pooling descriptors across multiple domain sizes. Meshroom is used with the `ultra` preset and the SfM matching is configured in exhaustive mode with the following parameters: distance ratio $= 0.75$, maximum reprojection error $= 1.0$ px. Meshroom is run directly on the native 16-bit EXR images, preserving the full radiometric precision of the acquisition and avoiding the degradation of sub-pixel feature localisation that can result from 8-bit quantisation. The resulting bundle adjustment achieves a root-mean-square reprojection error of 0.095 px, used here as an internal consistency metric.

B Segmentation & Alignment Specifications

B.1 Canny Explanation

To establish accurate 2D–3D registration, we employ a contour-based alignment protocol that refines initial segmentation boundaries through optimised edge detection.

Contour extraction. For each view i, target contour points $\mathcal{C}_i$ are extracted from the single-light images captured under rotation-invariant ambient illumination. For highly specular objects, images acquired with sublimating opacifying spray are used instead to ensure clean silhouettes. For the KNIFE, we obtain the best alignment by jointly registering the scan to the single-light and multi-light images. A search band of $m = 20$ pixels is constructed around the initial Segment Anything Model (SAM) mask boundary by morphological dilation and erosion:

$$\mathcal{M}_i = \mathrm{dilate}(S_i, m) \setminus \mathrm{erode}(S_i, m), \tag{1}$$

and the target contour is defined as $\mathcal{C}_i = \mathrm{Canny}(I_i) \cap \mathcal{M}_i$, retaining only edge pixels that fall within this band.

Registration objective. The alignment minimises the sum of nearest-neighbour distances from the target Canny contour points to the projected silhouette vertices of the ground-truth mesh:

$$\min_{R,\,\mathbf{t},\,s} \sum_{i=1}^{N} \sum_{\mathbf{c}\in\mathcal{C}_i} \min_{j\in\mathcal{S}(M,\,C_i)} \left\| \mathbf{c} - \pi(R\,(s\,\mathbf{x}_j) + \mathbf{t},\; K_i) \right\|, \tag{2}$$

where $\mathbf{x}_j$ are vertices of the ground-truth mesh and $\mathcal{S}(M, C_i)$ is the set of silhouette vertices visible from camera C_i. The set $\mathcal{S}$ includes edges whose adjacent faces have opposite front/back visibility, filtered by rasterisation to retain only occluding contour points. The perspective projection $\pi(\cdot, K_i)$ accounts for radial distortion, and the nearest projected silhouette vertex of each contour point is found with a KD-tree. The solver uses the Trust Region Reflective method with a soft-ℓ_1 loss for robustness to spurious Canny edges, in a coarse-to-fine scheme with outlier trimming.[5]

Threshold optimisation. The Canny edge detection thresholds are selected per object from a grid of 20 configurations: $\tau_\ell \in \{5, 10, 20, 50\}$ and $\tau_h \in \{20, 30, 40, 60, 80, 120\}$ with the constraint $\tau_\ell < \tau_h$, a Sobel kernel size of 15, and a search margin of 20 px. For each configuration, a registration is performed and the result is evaluated by rasterising the ground-truth mesh at the registered pose to produce a reference silhouette S for each view.

For a given tolerance ε (in pixels), the evaluation metrics are aggregated across views:

- **Silhouette coverage** $T_R = |\{s \in S : d(s, C) \le \varepsilon\}| \,/\, |S|$: fraction of the GT silhouette covered by Canny edges.
- **Precision after refinement** $P_{\text{trim}} = |\{c \in C : d(c, S) \le \varepsilon\}| \,/\, |C|$: fraction of refined Canny points that lie on the GT silhouette.
- $F_{0.5}$ **score**:

$$F_{0.5} = \frac{1.25 \cdot T_R \cdot P_{\text{trim}}}{0.25 \cdot P_{\text{trim}} + T_R}, \tag{3}$$

  which weights precision more than recall, favouring clean contours over high coverage.

The configuration maximising $F_{0.5}$ at $\varepsilon = 2$ px is retained for each object.

B.2 Limitations of the IoU Score for Alignment Validation

EvalMVX [21] uses IoU between projected ground-truth masks and segmented image masks as its alignment-quality metric. Our analysis suggests that such global overlap scores can remain very high despite visible contour errors. On two objects from our dataset, ROCK and WARRIOR HEAD, SAM masks reach IoU values above 0.99 against manually segmented reference masks. However, their contours differ by 4.80 and 4.51 pixels on average, respectively, boundary shifts that are poorly captured by IoU. See Figure 9.

As shown in Table 4, incorporating Canny edge detection into our segmentation pipeline significantly enhances contour extraction compared to standard SAM masks. This refinement reduces the average contour distance from the 4–5 pixel range to between 1 and 2 pixels.

[5] Implementation details are available in the released registration code.

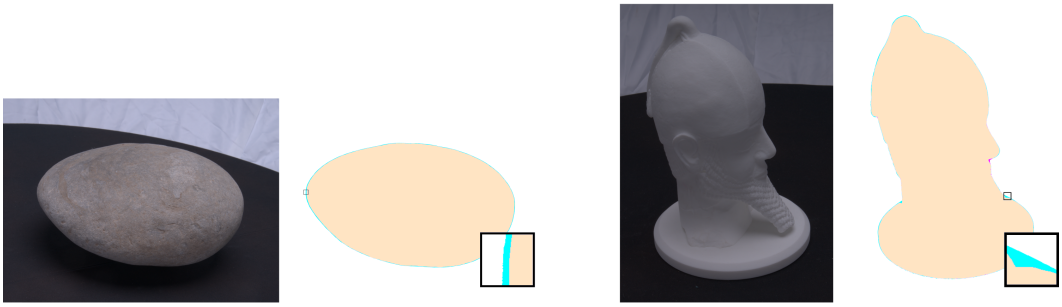


Figure 9: IoU limitations in assessing boundary precision. Visualisation of the discrepancy between SAM (magenta) and ground-truth (cyan) masks, with their intersection in orange. The boundary analysis reveals systematic contour shifts (4.80 and 4.51 px on average) that are critical for 2D–3D registration accuracy.

Table 4: **Contour refinement.** The proposed Canny-based refinement significantly reduces the mean contour distance (px).

	SAM		SAM + Canny (Ours)	
Object	**IoU ↑**	**Dist. (px) ↓**	**IoU ↑**	**Dist. (px) ↓**
ROCK	0.9918	4.80	0.9972	1.28
WARRIOR HEAD	0.9913	4.51	0.9984	1.37

B.3 OpenSubstance Dataset Limitations

While the OpenSubstance [20] dataset provides a vast collection of objects, a significant portion of the library is unsuitable for rigorous qualitative or quantitative 3D reconstruction analysis. Figure 10 illustrates a representative subset of these problematic cases. The first two examples possess ground-truth models, yet their geometric characteristics pose significant challenges for COLMAP; this leads to substantial registration errors between the estimated camera poses and the ground-truth geometry. The two remaining examples have highly complex geometries that could not be captured reliably by the scanning process and therefore do not have ground-truth data. This information was shared by the dataset authors.



Figure 10: Sample of OpenSubstance objects that are unsuitable for rigorous 3D reconstruction analysis.

C Evaluation of the View-Invariant Lighting Setup

To validate the efficacy of our single-light acquisition regime, we conducted a two-stage experimental analysis focusing on the temporal stability and the rotational invariance of the single-light setup. As detailed in the main manuscript, this setup utilises a top-mounted softbox flash and a reflecting cone encircling the turntable. When perfectly aligned, this configuration is designed to produce a view-invariant lighting environment optimised for multi-view reconstruction.

Photometric stability. We first assessed the repeatability of the flash intensity by capturing 50 successive images of a static object from a fixed viewpoint. By computing the standard deviation for each pixel across the sequence, we observed an average standard deviation of 0.005 on a normalised intensity scale of $[0, 1]$. This low variance confirms the photometric stability of the flash and its suitability for consistent data acquisition.

Rotational invariance. To evaluate the view-invariant nature of the setup, we performed a symmetry analysis using a high-fidelity white sphere (25 cm in diameter) positioned at the centre of the turntable. Following camera calibration and the geometric alignment of a 3D spherical model with the image data, we identified a series of isophotes—contours of equal elevation—across the sphere's surface. Given the rotational symmetry of the capture environment, the radiance values along these isophotes should ideally remain constant regardless of the turntable's rotation. We extracted intensity data from 41 viewpoints across multiple isophotes spanning the height of the sphere. The resulting distribution, visualised in the box-and-whisker plots in Figure 11, demonstrates the precision of our lighting invariance. The observed standard deviations range between 0.0054 and 0.0267 (mean 0.0141). Considering that a portion of this variance is attributable to the inherent flash instability identified previously, these low values confirm that the lighting remains largely independent of rotation.

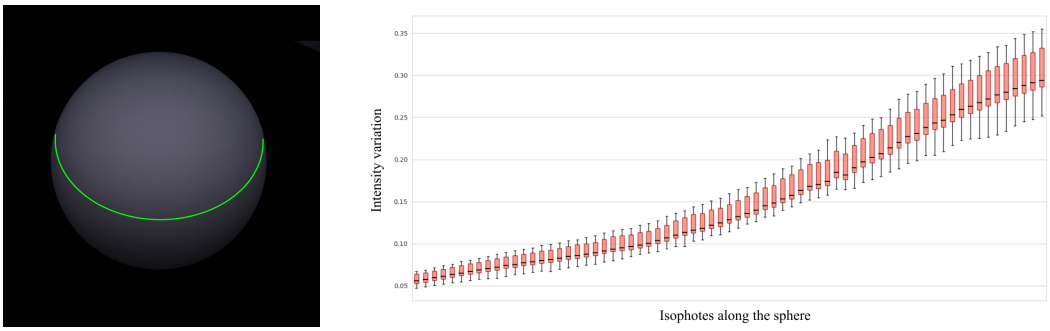


Figure 11: (Left) A visualisation of a single isophote (green) on the calibration sphere. (Right) Distribution of pixel intensity variations for each isophote across the sphere's height. The standard deviation ranges from 0.0054 to 0.0267, with an average of 0.0141, confirming consistent illumination across viewpoints.

D Light Calibration

D.1 Multi-light Calibration

Multi-light methods are evaluated on our dataset using normals predicted by UniMS-PS [25], a neural photometric-stereo method commonly used in recent multi-light reconstruction work and not requiring calibrated light sources. However, many calibrated photometric-stereo methods require light positions, directions, or intensities as input. To support such methods, we provide additional images for light calibration. Following standard sphere-based calibration procedures, we acquire multi-view multi-light sequences of two registered spheres placed in the object coordinate frame. The first is a matte sphere of diameter 60.5 mm, which can be used to estimate photometric response and intensity-related quantities. The second is a glossy sphere of diameter 60.2 mm, whose highlights provide constraints on light directions. Because the sphere poses are known from our image-to-geometry alignment, we used them to estimate the light parameters based on two models: an anisotropic model and an isotropic model. These models are provided alongside the data. Further details regarding the light calibration are available on the project page.

D.2 Photometric Calibration

To support photometric accuracy and texture consistency, we include captures of a professional colour calibration target (Calibrite ColorChecker Classic) for each LED and the single-light scenario. These reference images allow for the correction of light colour temperatures in photometric stereo pipelines and ensure that the final reconstructed meshes feature accurate and consistent textures.

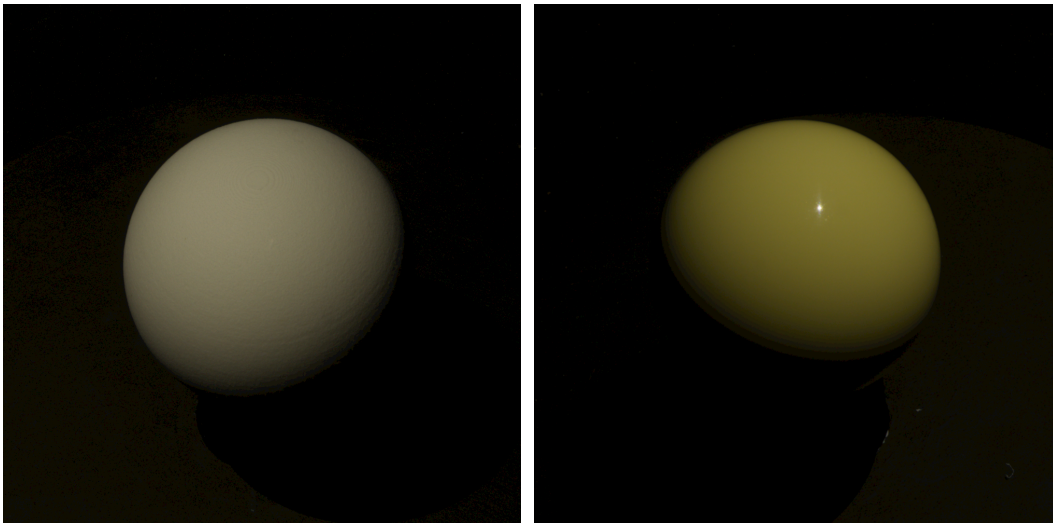


Figure 12: Light-calibration spheres. (Left) Matte sphere. (Right) Glossy sphere.

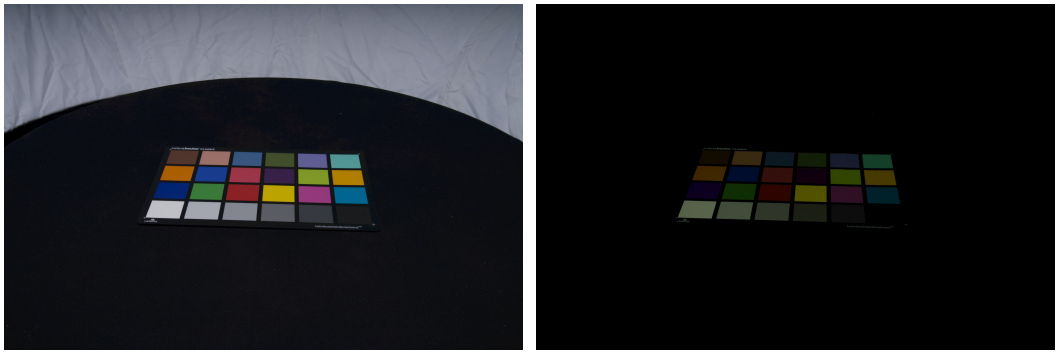


Figure 13: Colour calibration board seen in the single-light regime (left), and in the multi-light regime with one LED turned on (right).

E Additional Information

E.1 Evaluation Protocol

Mesh cleanup. Reconstructed and ground-truth meshes are filtered in the same way, by removing vertices that are not visible from any camera view or that project outside the object masks. Mounting bases are removed from the ground-truth meshes and from the masks of the objects, and unreliable regions of the ground-truth meshes (e.g., the base of the CANDLE or sharp rims) are excluded using per-object 3D masks. The ground truth of WOOD and SNEAKERS is open (no underside, no interior); reconstructed surfaces seen through these openings, in per-view masks, are removed.

View subsets. Subsets of N views are selected by k-means clustering of the 3D camera centres (N clusters, fixed seed), keeping the camera closest to each centroid; all methods using a sequence share its subsets.

All quantitative evaluations are performed by comparing the reconstructed mesh $\mathcal{P}$ against the ground-truth mesh $\mathcal{G}$. Following standard benchmarking protocols, we report precision, recall, F-score and Chamfer distance.

The F-score is defined as the harmonic mean of precision $P(\epsilon)$ and recall $R(\epsilon)$ for a given distance threshold ϵ:

$$F(\epsilon) = \frac{2P(\epsilon)R(\epsilon)}{P(\epsilon) + R(\epsilon)}, \tag{4}$$

where precision and recall represent the fraction of points whose distance to the reference mesh is below ϵ:

$$P(\epsilon) = \frac{1}{|\mathcal{P}|} \sum_{\hat{\mathbf{x}} \in \mathcal{P}} [d_{\hat{\mathbf{x}} \to \mathcal{G}} < \epsilon] \quad \text{and} \quad R(\epsilon) = \frac{1}{|\mathcal{G}|} \sum_{\mathbf{x} \in \mathcal{G}} [d_{\mathbf{x} \to \mathcal{P}} < \epsilon]. \tag{5}$$

Here, $[\cdot]$ denotes the Iverson bracket.

To assess the proximity between the two surfaces, we define the distance from a reconstructed point $\hat{\mathbf{x}} \in \mathcal{P}$ to the ground-truth mesh $\mathcal{G}$ as:

$$d_{\hat{\mathbf{x}} \to \mathcal{G}} = \min_{\mathbf{x} \in \mathcal{G}} \|\hat{\mathbf{x}} - \mathbf{x}\|. \tag{6}$$

The distance from the ground truth to the reconstruction, $d_{\mathbf{x} \to \mathcal{P}}$, is defined analogously. These measurements are accumulated over the entire surface to compute the Chamfer distance (CD):

$$CD = \frac{1}{2} \left(\frac{1}{|\mathcal{P}|} \sum_{\hat{\mathbf{x}} \in \mathcal{P}} d_{\hat{\mathbf{x}} \to \mathcal{G}} + \frac{1}{|\mathcal{G}|} \sum_{\mathbf{x} \in \mathcal{G}} d_{\mathbf{x} \to \mathcal{P}} \right). \tag{7}$$

Distances larger than 4 mm are discarded from the Chamfer distance; precision and recall use all points.

To ensure the reliability of these metrics, it is essential that the point densities of the reconstructed mesh $\mathcal{P}$ and the ground-truth mesh $\mathcal{G}$ are comparable. A significant discrepancy in density can bias both the Chamfer distance and the F-score, as sparsely populated regions may disproportionately affect the nearest-neighbour search. To mitigate this, we implement a consistent resampling pipeline applied to both the ground-truth and reconstructed meshes prior to evaluation. Each mesh is first converted into a dense point cloud by sampling each face on a regular grid of 0.05 mm spacing. This is followed by a spatial downsampling filter with a 0.05 mm radius to eliminate local clusters and enforce a uniform distribution. This procedure yields, for every object, two point clouds (ground truth and reconstruction) with the same inter-point distance of about 0.05 mm, providing a standardised and fair baseline for the final computation of the Chamfer distance and F-score.

The precision, recall and F-score curves and the radar chart pool the point distances of all objects, so that larger objects weigh more; a failed reconstruction keeps its ground-truth points in the pool, all counted as missed, which lowers recall and F-score but leaves precision unchanged.

Curvature. High-curvature regions are identified on the ground truth. First, each ground-truth mesh is densified by splitting all edges longer than 0.15 mm at their midpoint, with the same threshold for every object. The densified mesh describes exactly the same surface as the original one, but with a fine and regular sampling. This step is needed because some ground-truth meshes, mainly those obtained by CT scanning, contain very elongated triangles near which curvature would otherwise be overestimated.

Then the curvature is computed at each vertex of the densified mesh with the Voronoi Covariance Measure (VCM) [40], a robust estimator based on the shape of the Voronoi cells of the surface samples. It depends on two radii: an offset radius R, which bounds the extent of each cell, and an integration radius r, which sets the size of the neighbourhood over which the curvature is measured. At each vertex, it gives the two principal curvatures, of which we keep the one of largest magnitude, k. Each point of the ground-truth point cloud then takes the curvature of its nearest vertex in the densified mesh.

Finally, points with $|k| > 0.75$ mm^{-1} form the high-curvature region and the remaining points the flat region. The values $R = 0.5$ mm, $r = 1$ mm and the threshold of 0.75 mm^{-1} were chosen after testing several values on objects of the dataset, as the setting that best separated the high-curvature regions. The precision, recall, F-score and Chamfer distance are computed separately on each region, which isolates the behaviour of each method on the most demanding parts of the surface.

E.2 Data Release and Format

For each object, the release provides the single-light images at four resolutions, d1 (full resolution, 9568×6376), d2, d4 and d8 (downsampled by factors of 2, 4 and 8 with area averaging), and the multi-light images (84 poses $\times$ 30 LEDs) at d2, d4 and d8. Images are stored as 16-bit half-float OpenEXR files, together with binary PNG object masks (at each lower resolution, the area average of the full-resolution mask thresholded at 128 of 255) and an AliceVision `sfm.json` file containing intrinsics and poses, with the centre of the top-left pixel at $(0, 0)$. For each evaluated object, the ground-truth kit contains the mesh, the point cloud sampled for evaluation, per-point curvature and visibility, the 3D challenge masks and the evaluation masks; the light calibration is shared by all objects. The release totals 680 GB and is described by a Croissant metadata file. Converters to the input formats of all evaluated methods, with their pixel-centre conventions, and download scripts are provided in the code repository.

E.3 Hardware and Acquisition Parameters

The single-light acquisition setup was controlled using a Godox DP1000III-V strobe with a maximum power of 1000 W, paired with a SmallRig RA-D120 softbox to ensure uniform light diffusion. All images were captured using a Sony α7R V camera. To maintain consistent depth of field across varying object geometries, a narrow aperture of $f/20$ and a sensitivity of ISO 100 were utilised for all captures.

The exposure settings were adjusted according to the acquisition paradigm as follows:

- **Single-light acquisition:** To effectively freeze the flash output and eliminate ambient light interference, the shutter speed was set to 1/250 s.
- **Multi-light acquisition:** To accommodate the LED intensities, a longer shutter speed of 1 s was employed.

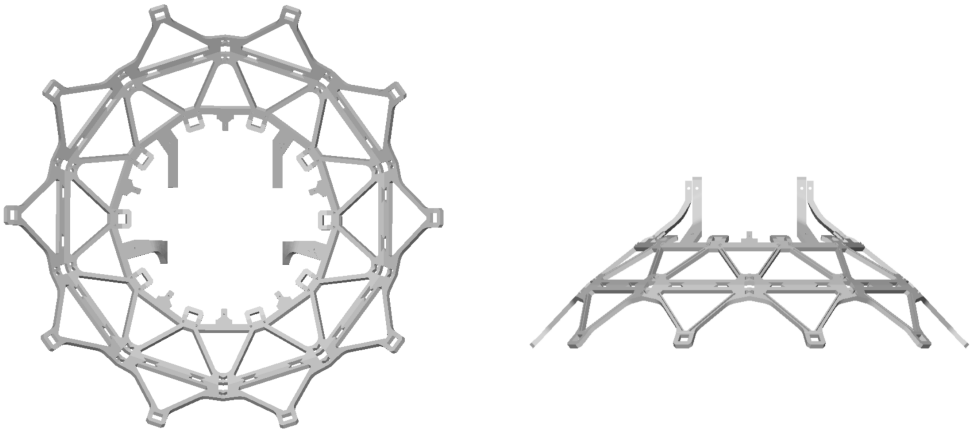


Figure 14: LED rig. Two views showing the spatial distribution of the LEDs around the camera.

E.4 Baseline Settings

Table 5 lists the evaluated methods and their main settings are given below; their exact versions are listed in the converters documentation of the dataset repository (footnote in Section 1).

Common settings. All methods use the calibrated cameras and the images of the evaluated resolution level or view subset. The sparse points required by COLMAP and PGSR are triangulated from these images only, with the calibrated poses kept fixed. NeuS2, Neuralangelo, SuperNormal and RNb-NeuS2 are normalised to the minimal sphere enclosing the visual hull of the object masks, enlarged by 5%, and their meshes are extracted by marching cubes on a 1024^3 grid at the zero level set. The number of iterations is adapted to the resolution level and the number of images, following

Table 5: **Evaluated methods.**

Method	Family
Meshroom [3]	Photogrammetry
RealityScan [2]	Photogrammetry
COLMAP [4, 22]	Classical MVS
PGSR [36]	3D Gaussian Splatting
NeuS2 [37]	Neural implicit
Neuralangelo [7]	Neural implicit
SuperNormal [8]	MVPS
RNb-NeuS2 [17]	MVPS
UniMS-PS [25]	Photometric stereo (normals)

the reference papers. Otherwise, each method uses the default configuration of its authors; the changes are listed below.

Meshroom. Release 2025.1.0.

RealityScan. Version 2.1.0 (Linux). Poses and calibration imported through XMP files and locked, without distortion and with the object masks. Default alignment, no image downscaling, feature detection set to high, with tie points computed from the images of the run only, followed by a high-detail reconstruction.

COLMAP. Version 3.12.6, default dense pipeline, with the stereo fusion restricted to the object masks.

PGSR. Our patched version uses the calibrated principal point, which the official code replaces by the image centre. The TSDF parameters are set to the scale of the objects (voxel 0.5 mm, truncation 2 mm).

NeuS2. Official DTU configuration.

Neuralangelo. Default configuration (500k iterations), at quarter resolution.

SuperNormal. Official configuration.

RNb-NeuS2. Default configuration.

UniMS-PS. Uncalibrated model, one inference per view on all 30 single-LED images at the evaluated resolution level, cropped to the bounding box of the object mask dilated by 5 pixels.

E.5 Additional Results

Table 6 reports the Chamfer distance and the F-score at 0.5 mm of every method on the 26 objects with ground truth, across all tested image resolutions, from d1 (full resolution) to d8 (one-eighth resolution).

Table 7 gives, for the representative method of each family, the number of objects reconstructed for each subset of input views and each image resolution, and Table 8 lists the failed reconstructions. These failures are real failures of the methods, not missing computations: COLMAP, PGSR and RealityScan start from a sparse model computed from the images of the evaluated configuration only, with a single attempt and default settings, and fail when this model is too sparse.

Across objects, view sparsity affects the methods as on the VASE (Figure 5): with 10 views, COLMAP, PGSR and RealityScan reconstruct 14, 13 and 24 of the 26 objects, and 4, 3 and 13 with 5 views, whereas NeuS2 and RNb-NeuS2 reconstruct all of them in every configuration. At one-eighth resolution, COLMAP, PGSR and RealityScan fail on 5, 4 and 3 objects, respectively.

Table 6: Chamfer distance (mm, Eq. 7) / F-score at 0.5 mm (%) per object, method and image resolution (d1 to d8: full to one-eighth resolution), with all 84 views; Meshroom at d4 only; methods in two sub-tables, one under the other. **Bold**: best per row over both sub-tables, underline: second best, for each metric separately (lowest Chamfer distance, highest F-score). –: failure of the method (no reconstruction, Table 8). *Mean*: point distances of all objects pooled, as if they formed a single mesh, so that larger objects weigh more; a failed reconstruction keeps its ground-truth points in the pool, all missed: they lower the F-score, and lying beyond the cutoff of Eq. 7 they do not enter the Chamfer distance, so Chamfer means over different numbers of reconstructed objects (*Reconstructed* row) are not directly comparable.

	RealityScan				Meshroom	COLMAP			PGSR		
Object	d1	d2	d4	d8	d4	d2	d4	d8	d2	d4	d8
ROCK	0.108 / 99.1	0.106 / 99.1	0.148 / 98.4	0.261 / 90.4	0.108 / 98.2	0.102 / 98.2	0.115 / 97.5	0.156 / 94.8	0.340 / 87.4	0.508 / 68.7	0.774 / 12.7
VASE	**0.135 / 95.9**	0.256 / 77.4	0.638 / 41.3	1.135 / 19.1	0.309 / 82.5	0.243 / 79.3	0.267 / 73.7	–	0.577 / 50.0	0.625 / 49.0	–
HAMMER HEAD	0.088 / 98.9	0.104 / 99.0	0.174 / 97.8	0.459 / 65.3	0.102 / 98.4	0.065 / 99.2	0.110 / 97.5	0.865 / 8.0	0.161 / 98.1	0.282 / 95.6	0.566 / 29.9
CANDLE	0.821 / 43.6	1.106 / 33.3	1.362 / 21.8	2.004 / 7.2	1.546 / 16.5	0.755 / 39.2	0.895 / 25.5	–	0.486 / 87.8	0.587 / 62.7	–
CAN	**0.149 / 96.9**	0.184 / 96.2	0.276 / 90.1	0.672 / 44.1	0.364 / 83.0	0.250 / 85.7	0.254 / 89.1	0.639 / 43.9	0.500 / 67.0	0.585 / 57.9	0.726 / 42.9
GLASS	1.525 / 20.7	1.456 / 17.7	1.759 / 7.9	1.999 / 4.7	2.535 / 0.3	1.224 / 24.8	1.345 / 21.8	1.415 / 12.9	1.289 / 26.6	1.287 / 25.6	1.356 / 20.5
PETANQUE BALL	0.158 / 99.0	0.210 / 97.2	0.378 / 79.2	0.993 / 9.1	0.232 / 94.4	0.145 / 97.5	0.361 / 63.9	1.422 / 0.2	0.294 / 97.1	0.375 / 81.3	0.599 / 24.5
JAM POT	0.964 / 39.8	1.018 / 36.8	1.141 / 33.1	1.373 / 25.1	1.551 / 19.2	0.973 / 33.9	0.910 / 40.0	1.021 / 30.9	0.973 / 21.4	1.254 / 12.8	1.669 / 4.7
COLANDER	0.408 / 69.8	0.415 / 70.0	0.731 / 38.0	0.929 / 30.5	0.460 / 65.1	0.622 / 54.5	0.551 / 61.5	0.620 / 53.2	0.394 / 73.2	**0.368 / 75.8**	0.387 / 72.5
WOOD	**0.121** / 97.4	0.166 / 94.9	0.277 / 85.6	0.463 / 67.7	0.227 / 89.4	0.171 / 92.8	0.243 / 88.7	0.359 / 79.6	0.302 / 83.7	0.318 / 83.2	0.401 / 73.4
PLATE	0.080 / 99.2	0.077 / **99.4**	**0.069** / 99.3	0.149 / 95.6	0.088 / 99.4	0.104 / 98.3	0.114 / 98.0	0.147 / 92.5	0.148 / 92.2	0.340 / 80.2	0.648 / 67.8
PAW PAD	0.316 / 81.5	0.514 / 68.7	1.384 / 19.2	–	1.928 / 3.9	0.991 / 3.8	0.886 / 12.0	–	0.824 / 48.8	0.905 / 42.9	1.231 / 16.3
MOKA POT	**0.127 / 95.8**	0.158 / 93.2	0.258 / 79.0	0.504 / 58.4	0.231 / 88.3	0.272 / 84.7	0.327 / 80.5	0.590 / 60.4	0.749 / 47.4	0.779 / 44.5	0.953 / 17.4
LEGO	0.463 / 67.1	0.428 / 67.9	0.407 / 68.1	0.496 / 60.3	0.479 / 64.1	0.499 / 67.0	0.452 / 69.4	0.674 / 42.4	0.334 / 78.0	0.406 / 75.7	0.539 / 58.1
SURGERY SCISSORS	0.318 / 84.3	0.473 / 60.2	1.000 / 22.6	–	–	0.270 / 88.9	0.502 / 55.1	–	0.163 / 92.5	0.209 / 92.6	–
KNIFE	0.131 / **99.5**	0.165 / 97.9	0.262 / 92.2	0.562 / 50.6	0.142 / 97.2	**0.089** / 99.2	0.112 / 98.5	0.508 / 64.1	0.204 / 97.4	0.313 / 90.4	0.603 / 25.5
SPIKES BALL	0.385 / 70.2	0.495 / 59.8	0.725 / 45.4	0.919 / 33.6	0.719 / 46.9	0.464 / 66.0	0.543 / 54.4	1.029 / 7.4	0.553 / 57.5	0.483 / 61.4	0.577 / 42.5
STRAW HAT	0.178 / **99.4**	0.199 / 98.0	0.245 / 91.9	0.330 / 76.9	0.201 / 96.9	0.169 / 98.5	0.200 / 96.4	0.253 / 91.0	0.283 / 87.1	0.307 / 84.5	0.404 / 65.5
SNEAKERS	0.162 / 96.2	0.160 / **97.3**	0.223 / 92.9	0.527 / 59.7	0.276 / 89.4	0.173 / 94.6	0.195 / 93.6	0.286 / 86.1	0.312 / 81.2	0.315 / 82.1	0.460 / 57.5
SURGERY SAW	**0.204** / 91.4	0.245 / 88.4	0.632 / 61.1	1.221 / 24.6	0.792 / 53.1	0.398 / 75.9	0.510 / 69.2	1.024 / 39.6	0.361 / 80.5	0.422 / 74.0	0.624 / 26.8
WHISK	0.301 / 88.9	0.468 / 50.9	1.433 / 12.0	–	1.488 / 14.7	0.397 / 66.1	0.598 / 33.0	–	0.211 / 91.3	0.255 / 89.8	–
WARRIOR HEAD	0.188 / **94.3**	0.246 / 88.9	0.386 / 73.9	0.628 / 48.7	0.328 / 83.0	0.338 / 78.6	0.308 / 83.5	0.382 / 76.1	0.488 / 69.6	0.515 / 68.3	0.670 / 34.3
PLANT	0.403 / **78.7**	0.579 / 59.5	0.690 / 47.2	0.982 / 29.3	0.664 / 56.9	0.657 / 52.3	0.736 / 44.8	0.970 / 27.6	0.619 / 50.9	0.622 / 52.2	0.712 / 31.8
GRID	**0.216** / 93.4	0.363 / 81.9	0.428 / 60.9	0.838 / 28.1	0.408 / 64.1	0.434 / 74.2	0.601 / 63.2	0.875 / 27.1	0.347 / 78.2	0.346 / 75.6	0.568 / 49.6
BOWL	**0.276 / 90.5**	0.484 / 73.4	1.024 / 34.2	1.791 / 6.6	1.022 / 37.9	0.364 / 79.1	0.729 / 45.7	0.927 / 15.8	1.134 / 24.5	1.175 / 20.3	1.190 / 14.8
SPINE	0.465 / 70.6	0.528 / 65.1	0.732 / 43.6	1.022 / 26.7	0.742 / 46.1	0.508 / 65.5	0.579 / 58.4	0.843 / 33.4	0.569 / 61.7	0.552 / 65.3	0.632 / 43.3
Mean	0.333 / 82.4	0.381 / 77.4	0.536 / 63.6	0.731 / 45.4	0.477 / 68.1	0.380 / 74.5	0.432 / 70.6	0.557 / 54.7	0.503 / 65.9	0.559 / 62.1	0.702 / 40.1
Reconstructed	26/26	26/26	26/26	23/26	25/26	26/26	26/26	21/26	26/26	26/26	22/26

	NeuS2			Neuralangelo	RNb-NeuS2			SuperNormal
Object	d2	d4	d8	d4	d2	d4	d8	d4
ROCK	0.103 / 99.6	0.085 / 99.8	0.098 / 99.8	0.099 / 98.4	0.068 / **100.0**	0.108 / 99.9	0.370 / 82.1	**0.063** / 99.8
VASE	0.271 / 67.0	0.268 / 66.9	0.302 / 65.4	0.276 / 81.8	0.610 / 60.7	0.495 / 67.7	0.529 / 68.4	0.947 / 42.9
HAMMER HEAD	0.105 / 99.1	0.096 / 99.8	0.105 / 99.7	**0.058 / 99.8**	0.078 / 99.2	0.081 / 99.2	0.160 / 97.2	0.147 / 97.5
CANDLE	0.253 / 86.6	0.257 / 86.0	0.263 / 86.3	**0.247 / 89.9**	0.304 / 81.0	0.292 / 81.8	0.365 / 75.1	1.482 / 0.8
CAN	0.420 / 78.8	0.388 / 79.9	0.412 / 79.7	0.307 / 83.6	0.515 / 68.0	0.528 / 67.1	0.708 / 53.1	0.563 / 51.7
GLASS	1.051 / 31.1	0.880 / 46.0	**0.834** / 46.6	0.837 / **47.8**	1.256 / 19.5	1.247 / 19.3	1.258 / 20.1	2.006 / 5.1
PETANQUE BALL	0.218 / 92.8	0.162 / 93.9	0.185 / 90.6	0.178 / 94.3	0.113 / 99.3	**0.074** / 99.6	0.088 / **99.7**	0.126 / 96.7
JAM POT	**0.383 / 77.4**	0.386 / 76.3	0.438 / 70.6	0.550 / 63.9	1.203 / 9.9	1.254 / 10.2	1.213 / 10.4	1.255 / 10.9
COLANDER	0.805 / 42.1	0.782 / 44.0	0.796 / 41.2	0.599 / 51.8	0.574 / 51.7	0.569 / 52.8	0.776 / 39.7	0.488 / 62.6
WOOD	0.269 / 87.3	0.272 / 87.0	0.308 / 84.9	0.225 / 89.7	0.185 / **97.5**	0.161 / 96.8	0.172 / 96.5	0.145 / 94.1
PLATE	0.106 / 98.0	0.133 / 96.5	0.154 / 96.1	0.138 / 95.5	1.793 / 9.7	1.650 / 9.7	1.635 / 10.9	1.783 / 10.0
PAW PAD	0.461 / 71.9	0.461 / 71.2	0.460 / 69.6	**0.257 / 85.9**	0.874 / 33.0	0.881 / 33.4	0.906 / 31.5	1.327 / 18.4
MOKA POT	0.485 / 68.0	0.415 / 73.7	0.357 / 79.0	0.354 / 74.1	0.451 / 71.6	0.300 / 81.8	0.268 / 84.0	0.486 / 63.5
LEGO	0.326 / 80.1	0.332 / 79.8	0.334 / 79.5	**0.294 / 81.8**	0.604 / 59.2	0.539 / 61.1	0.613 / 58.2	0.723 / 56.9
SURGERY SCISSORS	0.080 / 99.3	0.089 / 99.0	0.083 / 99.1	1.691 / 8.5	**0.066 / 100.0**	0.073 / 100.0	0.086 / 99.8	0.084 / 99.3
KNIFE	0.134 / 98.4	0.141 / 97.7	0.141 / 98.4	0.278 / 89.3	0.231 / 87.5	0.198 / 91.2	0.216 / 89.7	0.380 / 77.1
SPIKES BALL	0.304 / 80.8	0.249 / 86.9	0.247 / 87.7	0.232 / 87.9	0.193 / 97.0	0.166 / 98.1	**0.137 / 98.3**	0.256 / 93.3
STRAW HAT	0.229 / 91.0	0.227 / 91.4	0.224 / 91.6	**0.167** / 97.9	0.283 / 88.6	0.234 / 91.3	0.292 / 85.5	0.283 / 88.2
SNEAKERS	0.265 / 82.8	0.288 / 81.4	0.282 / 81.2	**0.151** / 91.4	0.279 / 82.8	0.262 / 84.4	0.269 / 84.3	0.285 / 82.8
SURGERY SAW	0.256 / 87.6	0.252 / 87.6	0.241 / 88.9	0.214 / **92.3**	0.313 / 80.9	0.300 / 83.2	0.338 / 79.1	0.444 / 67.3
WHISK	0.126 / 96.2	0.124 / 97.8	0.134 / 95.6	–	0.084 / 98.9	**0.065 / 99.2**	0.109 / 98.5	0.097 / 98.1
WARRIOR HEAD	0.472 / 72.3	0.485 / 71.4	0.471 / 72.8	**0.175** / 94.1	0.277 / 86.5	0.283 / 86.4	0.284 / 86.6	0.373 / 78.4
PLANT	0.517 / 68.8	0.533 / 67.4	0.545 / 66.9	0.418 / 74.9	0.531 / 64.5	**0.395** / 76.0	0.422 / 73.8	0.462 / 68.5
GRID	0.219 / **93.5**	0.222 / 93.1	0.235 / 91.6	1.011 / 25.1	0.248 / 86.6	0.224 / 89.2	0.260 / 86.1	0.291 / 83.8
BOWL	0.546 / 50.5	0.591 / 48.6	0.560 / 51.3	0.824 / 44.3	0.864 / 34.0	0.927 / 35.7	0.999 / 31.9	1.165 / 24.5
SPINE	0.437 / 71.0	0.434 / 71.7	0.447 / 70.8	0.368 / 80.8	**0.202 / 91.2**	0.228 / 90.1	0.296 / 85.2	0.404 / 77.4
Mean	0.393 / 75.0	0.388 / 75.5	0.391 / 75.3	0.353 / 77.9	0.501 / 64.0	0.469 / 66.7	0.540 / 62.4	0.571 / 60.8
Reconstructed	26/26	26/26	26/26	25/26	26/26	26/26	26/26	26/26

Table 7: Number of objects reconstructed (out of 26) per method, for each subset of input views at quarter resolution (d4) and for each image resolution with all 84 views. A missing reconstruction is a failure of the method (Table 8). –: configuration not run; at full resolution, the methods other than RealityScan fail for lack of GPU memory (Section 3.2).

	Views (d4)					Resolution (84 views)			
Method	84	40	20	10	5	d1	d2	d4	d8
RealityScan	26	26	25	24	13	26	26	26	23
COLMAP	26	24	21	14	4	–	26	26	21
PGSR	26	24	21	13	3	–	26	26	22
NeuS2	26	26	26	26	26	–	26	26	26
RNb-NeuS2	26	26	26	26	26	–	26	26	26

Table 8: Failed reconstructions per configuration. COLMAP and PGSR fail when structure from motion on the images of the configuration alone yields no usable sparse model (single attempt, default settings, no fallback); the other failures are runs that produced no mesh. Failed objects are excluded from all aggregated scores and declared as $n/26$ in the captions.

Configuration	Failed	Objects
Meshroom d4	1/26	SURGERY SCISSORS
RealityScan d4, 20 views	1/26	SURGERY SCISSORS
RealityScan d4, 10 views	2/26	SURGERY SCISSORS, WHISK
RealityScan d4, 5 views	13/26	CANDLE, GLASS, PETANQUE BALL, MOKA POT, LEGO, SURGERY SCISSORS, SPIKES BALL, SURGERY SAW, WHISK, PLANT, GRID, BOWL, SPINE
RealityScan d8	3/26	PAW PAD, SURGERY SCISSORS, WHISK
COLMAP d4, 40 views	2/26	PAW PAD, SURGERY SCISSORS
COLMAP d4, 20 views	5/26	CANDLE, PAW PAD, SURGERY SCISSORS, SPIKES BALL, WHISK
COLMAP d4, 10 views	12/26	CANDLE, CAN, GLASS, PETANQUE BALL, COLANDER, PAW PAD, SURGERY SCISSORS, SPIKES BALL, SURGERY SAW, WHISK, BOWL, SPINE
COLMAP d4, 5 views	22/26	VASE, HAMMER HEAD, CANDLE, CAN, GLASS, PETANQUE BALL, JAM POT, COLANDER, PAW PAD, MOKA POT, LEGO, SURGERY SCISSORS, KNIFE, SPIKES BALL, STRAW HAT, SURGERY SAW, WHISK, WARRIOR HEAD, PLANT, GRID, BOWL, SPINE
COLMAP d8	5/26	VASE, CANDLE, PAW PAD, SURGERY SCISSORS, WHISK
PGSR d4, 40 views	2/26	PAW PAD, SURGERY SCISSORS
PGSR d4, 20 views	5/26	CANDLE, GLASS, PAW PAD, SURGERY SCISSORS, WHISK
PGSR d4, 10 views	13/26	CANDLE, CAN, GLASS, PETANQUE BALL, COLANDER, PAW PAD, SURGERY SCISSORS, SPIKES BALL, SURGERY SAW, WHISK, GRID, BOWL, SPINE
PGSR d4, 5 views	23/26	VASE, HAMMER HEAD, CANDLE, CAN, GLASS, PETANQUE BALL, JAM POT, COLANDER, PAW PAD, MOKA POT, LEGO, SURGERY SCISSORS, KNIFE, SPIKES BALL, STRAW HAT, SNEAKERS, SURGERY SAW, WHISK, WARRIOR HEAD, PLANT, GRID, BOWL, SPINE
PGSR d8	4/26	VASE, CANDLE, SURGERY SCISSORS, WHISK
Neuralangelo d4	1/26	WHISK

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions detailed in the abstract and introduction are directly supported by the dataset description and the experimental results provided in Sections 2 and 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and future work are discussed in the conclusion (Section 4).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Main settings of the methods are given in Section E.4 and their exact versions in the documentation of the dataset repository; the evaluation code is linked in Section 3.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset and the code are publicly available (footnotes in Sections 1, 2 and 3.2).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We do not have any training/test details on this dataset, but for the calibration process, we specify the operations performed and the parameters chosen (all in the appendix).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper does not report error bars or formal statistical-significance tests for the baseline reconstruction methods, which are presented as a first cross-paradigm evaluation. However, it does report quantitative accuracy statistics for the reference data.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: This paper focuses primarily on the introduction and rigorous validation of the dataset rather than the performance of specific reconstruction methods. Our objective is to establish the benchmark's quality and reliability; consequently, we prioritise the analysis of dataset fidelity over a detailed breakdown of the computational resources used for the experimental baseline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: This dataset focuses solely on technical contributions and benchmarking methodology without discussing broader societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset is released under CC BY 4.0 and the code under MIT; the licences are stated in the dataset, in the GitHub repositories and on the project page.

    Guidelines: The dataset and the code released on GitHub include the licence and the citation of the related paper.

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Croissant metadata and a README are provided; the data are released under CC BY 4.0 and the code under MIT.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.