

Mind the Domain Gap: Benchmarking Structure-from-Motion Methods on Paired Industrial Physical-Digital Twin Datasets to Enable Radiance Field Rendering

Sunil Choudhary^{1,2} , Mickyas Tamiru Asfaw¹ , Aristide Laignel¹ ,
Nicolas Ragot¹ , Vincent Havard¹ , Kristian Fenech³ , and Yohan
Dupuis⁴ 

¹ CESI, CESI LINEACT, Rouen, France

² Arts et Metiers Institute of Technology, Paris, France

³ Eötvös Loránd University, Budapest, Hungary

⁴ CESI, CESI LINEACT, Paris, France

Abstract. As industry transitions toward fully digitalized factories, mirroring physical environments in digital-twins has become essential for robotics, monitoring, and process automation, a capability that fundamentally relies on 6D camera pose estimation and 3D scene reconstruction. We propose an evaluation methodology to benchmark these tasks in paired physical and digital environments. Real-world acquisitions are performed on a diverse set of industrial objects - including a mobile robot, a robotic arm, a bike, and a bench drill - captured with a stereo camera, with accurate Ground-Truth (GT) poses provided by a Motion Capture (MoCap) system. To complement these acquisitions, we generate Digital-Twin (DT) counterparts by replaying the same camera trajectories in a similar Unity-based environment using the LOOPER tool. The resulting paired Physical-Twin (PT) and DT sequences share the same camera poses, enabling controlled analysis of the gap between real and simulated domains. We apply the methodology to evaluate several Structure-from-Motion (SfM) pipelines, assessing their pose accuracy. The resulting benchmark highlights the influence of image realism, object shape, and domain shift on pose estimation performance, providing a practical testbed for industrial vision research and supporting future work in 6D camera pose estimation, digital-twins, and 3D scene reconstruction. **Project page:** To access datasets, code, and related papers https://lineact-valo.gitlab.io/fusion_project/dataset_tools/

Keywords: Industrial Dataset · Digital-Twin · Physical-Twin · Structure-from-Motion · Motion Capture Ground Truth.

1 Introduction

Digital-Twins (DT) have emerged as a foundational technology in the Industry 4.0 paradigm, bridging the physical and digital domains to enable real-time

simulation, monitoring, and optimization of manufacturing operations [1]. To accurately align observations acquired in the physical world with their corresponding virtual models, robust camera localization and spatial understanding are critical prerequisites. Consequently, estimating accurate 6D camera pose and reconstructing 3D scene structure have become core capabilities in industrial computer vision [2]. These tasks are essential not only for digital twinning but also for applications such as robotic manipulation, inspection, autonomous navigation, and human-robot collaboration [3].

Among these capabilities, Structure-from-Motion (SfM) plays a particularly central role: its output - a set of camera poses together with a sparse 3D scene representation - serves as the standard input to modern 3D Radiance Field applications such as 3D Gaussian Splatting (3DGS) [4]. The accuracy of the SfM stage therefore directly conditions the fidelity of any downstream photorealistic reconstruction. Despite substantial recent progress in both visual localization and SfM pipelines [5], evaluating these methods under realistic industrial conditions remains highly challenging. Existing efforts such as [6] have begun to benchmark SfM on industrial-style scenes, but they typically rely on a single acquisition modality and do not provide paired real-simulated counterparts, leaving the influence of the synthetic-to-real domain gap largely unexplored.

Synthetic data generation has become increasingly important for scalable training and evaluation in computer vision. Platforms such as Unity enable richly annotated virtual environments with systematic variation of acquisition parameters [7]. Yet despite strong control and annotation quality, synthetic datasets often diverge from real acquisitions in appearance, sensor noise, lighting, and scene complexity. Whether methods that succeed in simulation generalize to physical industrial conditions therefore remains an open question.

We address these limitations by proposing an *evaluation methodology* for SfM on industrial objects: paired Physical-Twin (PT) and Digital-Twin (DT) sequences sharing identical camera trajectories. We instantiate this dataset by capturing image sequences of four industrial objects in the physical world, with a Motion Capture (MoCap) system providing GT poses, and recreating the same scenes in Unity by replaying the original trajectories using *LOOPER* (Light Object and event Oriented Package recordER), a framework for synthetic dataset generation with configurable sensors, integrated in Unity and editable for contextualizing data [8]. Replaying real sequences with same camera poses enables controlled analysis of the physical-to-digital gap while preserving realistic trajectory structures.

In particular, this work investigates two research questions:

- **RQ1:** To what extent does the physical-to-digital domain gap affect the camera pose accuracy of state-of-the-art SfM pipelines, and which scene properties (image realism, object shape, and texture richness) most strongly drive this discrepancy?
- **RQ2:** Is the ranking of SfM pipelines stable across industrial objects, or does the best-performing pipeline vary from one object to another?

The main contributions of this work are as follows.

- We propose an evaluation methodology for benchmarking SfM on industrial objects, explicitly designed to characterize the gap between PT and DT.
- We present the dataset generation methodology with a paired dataset of four different industrial objects: a PT set with precise GT camera poses acquired via a MoCap system and a DT counterpart generated by replaying matched real-world camera trajectories in Unity using the *LOOPER* tool.
- We provide a benchmark of state-of-the-art SfM pipelines across both PT and DT domains, analyzing the impact of physical-to-digital domain shifts.

2 Related Work

Evaluation frameworks and benchmark studies for 3D reconstruction. Standardized benchmarks have driven much of the progress in SfM and multi-view 3D reconstruction. Tanks and Temples [6] introduced large-scale outdoor and indoor scenes with laser-scanned GT, and ETH3D [9] provides multi-view stereo benchmarks at varying scales and resolutions. The Image Matching Challenge [10] systematically evaluates the components of SfM pipelines - feature detectors, descriptors, and matchers - across diverse imaging conditions. While these benchmarks have been instrumental for general 3D vision research, they are predominantly composed of outdoor landmarks, cultural-heritage sites, or controlled tabletop scenes, and none of them characterize SfM behavior on industrial objects or evaluate them on paired real-and-simulated counterparts.

Object pose datasets and industrial-oriented evaluation. A complementary line of work targets object-level 6D pose estimation. The BOP challenge [11] unifies several benchmarks under a common evaluation protocol, and datasets such as T-LESS [12] and YCB-Video [13] have advanced the study of texture-less, symmetric, and cluttered objects. These efforts, however, evaluate the localization of a known model within a scene, rather than the recovery of full camera trajectories. Consequently, they do not address the questions investigated in this work - how SfM pipelines behave on industrial datasets, and how their performance shifts between paired PT and DT acquisitions of the identical camera trajectory.

Synthetic data and PT-DT pairing. Synthetic dataset generation has become a standard tool for scalable evaluation, with established pipelines for scene randomization and rendering [7], and pre-rendered synthetic acquisitions under degraded conditions have also been used to support tasks such as 6D pose refinement using Gaussian Splatting [14]. However, most existing efforts treat the synthetic and the real domains as separate evaluation settings rather than as paired counterparts of the same trajectory. As a result, the influence of the physical-to-digital gap is typically assessed at the aggregate distribution level, not under matched viewpoints. Our work differs in this respect: by replaying real MoCap-tracked trajectories inside a Unity-based digital-twin, we obtain pairs of sequences whose only systematic difference lies in the rendering domain itself, enabling a controlled, per-view analysis of the gap.

SfM pipelines for evaluation. For benchmarking, four SfM pipelines are used. COLMAP [15] is the canonical incremental pipeline and remains the de facto reference for camera registration accuracy. GLOMAP [16] revisits global SfM for improved scalability, FastMap [17] targets efficient reconstruction at large scale, and InstantSfM [18] pushes this direction further with a fully parallel and sparse design oriented toward fast pose recovery. While each of these methods has been individually validated on the general benchmarks discussed above, their comparative behavior on industrial datasets and across paired PT and DT domains used as input to radiance field reconstruction has not been systematically characterized, which is the primary aim of this work.

3 Benchmarking Methodology

Role of SfM in radiance field rendering. Structure-from-Motion is a decisive stage for radiance field photorealistic reconstruction. Radiance field methods such as 3D Gaussian Splatting (3DGS) [4] depend directly on the camera poses and sparse scene geometry produced by SfM, and pose inaccuracies propagate as photometric errors and visible artifacts in the rendered output. Two outputs of the SfM stage are particularly important when targeting downstream Gaussian Splatting (GS) training: (i) the size of the sparse point cloud, which seeds the initial Gaussian positions, and (ii) the time required to obtain the sparse reconstruction, which determines whether the SfM stage is a bottleneck in the GS pipeline. As illustrated in Fig. 1, our benchmark therefore evaluates SfM in the specific context of its use as the front-end of neural rendering, and is designed to expose these two outputs alongside pose accuracy.

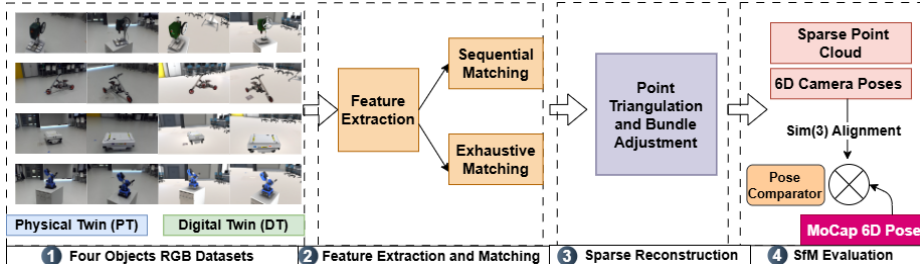


Fig. 1. Pipeline for evaluating SfM methods on paired Physical-Twin (PT) and Digital-Twin (DT) datasets.

Choice of SfM pipelines. We evaluate four pipelines whose selection is motivated by their growing adoption as the front-end of neural rendering applications: COLMAP [15], GLOMAP [16], FastMap [17], and InstantSfM [18]. These pipelines differ along several axes that are useful for interpreting their comparative behavior. COLMAP follows the classical *incremental* paradigm, sequentially registering cameras and refining the reconstruction through repeated bundle adjustment, which yields high accuracy at the cost of runtime. GLOMAP adopts a

global formulation in which camera poses are estimated jointly through rotation and translation averaging, trading some incremental robustness for substantially better scalability. FastMap and InstantSfM represent a more recent generation of *efficiency-oriented* designs: FastMap accelerates large-scale reconstruction by streamlining the matching-to-mapping flow, whereas InstantSfM leverages a fully sparse and parallel formulation to drastically reduce reconstruction time. These design differences directly affect the two GS-relevant outputs identified above, which our benchmark is designed to evaluate.

The benchmark is built on our paired PT-DT dataset of four industrial objects as shown in multiple stages in Fig. 1. The PT-DT datasets are shown in stage 1; then in stage 2, we first extract local image features using SIFT, and then perform feature matching using two complementary strategies. *Exhaustive* matching is used to maximize correspondence recovery and provide strong connectivity across views, whereas *sequential* matching is also considered to better reflect the temporal continuity of the acquisition and to reduce incorrect associations in repetitive, symmetric, or weakly textured industrial objects. The resulting matches are passed for point triangulation and bundle adjustment which is common in all four SfM pipelines in stage 3, and each pipeline produces a set of registered camera poses together with a sparse 3D point cloud. In stage 4, we evaluate the estimated 6D camera poses against Ground-Truth (GT) trajectories after Sim(3) - alignment to the MoCap reference frame, and additionally report the sparse point-cloud size and the reconstruction runtime as the two outputs most directly relevant for radiance fields training.

4 Dataset Overview

This benchmark dataset comprises two complementary paired components: a PT dataset and a DT dataset (blue for Physical-Twin, green for Digital-Twin in Fig. 1,2,3). Both subsets represent similar industrial objects with identical camera motions to enable direct analysis of the sim-to-real domain gap. The PT dataset consists of real-world acquisitions of diverse industrial objects, such as a mobile robot MiR (515 images), a robotic arm Niryo (220 images), a bike (515 images) and a bench drill (275 images) captured with a ZED Mini stereo camera. A MoCap system continuously records the camera trajectory, providing accurate 6D GT poses. The DT dataset mirrors this setup in a virtual environment. Industrial objects, lighting, and scene layouts are simulated in the Unity game engine, and the physical camera trajectories are replayed using *LOOPER* tool. Dataset codes in tables: BD=Bench Drill, BK=Bike, MR=Mir, NR=Niryo.

PT Dataset Acquisition This setup is specifically designed to capture realistic industrial environments under controlled, measurable conditions. The physical-twin is based on the CESI LINEACT Industry of the Future platform, a facility equipped with industrial objects and robotic arms. To obtain highly precise 6D poses (translation and orientation), we utilize a MoCap system comprising 16 tracking cameras distributed throughout the laboratory. By rigidly attaching reflective markers to the camera, we track its pose and reference its optical center

within the MoCap coordinate system via an initial extrinsic calibration. During data collection, the operator navigates the camera around the targeted objects from multiple viewpoints to acquire comprehensive scene coverage in a stop-and-go motion to have the MoCap have a stable and accurate GT camera pose. Object poses are obtained by recording specific keypoints using a dedicated tool equipped with MoCap markers. For each object, five reference points are acquired: four located at the symmetric corners of the object’s base and one at its top to capture the object height. The 3D center of the object is then computed from these five points, providing an accurate and reproducible pose reference within the MoCap coordinate system. As part of this process, images and their corresponding MoCap GT are recorded only when the camera markers remain visible to at least two MoCap cameras simultaneously. The final PT dataset has multimodality including full-HD left and right RGB images, depth maps, camera’s internal tracking pose with intrinsics (6-DoF visual-inertial stereo SLAM with sensor fusion; IMU: accelerometer (m/s^2) and gyroscope (rad/s) at 800 Hz data rate, pose drift of $\pm 1.0\%$ translation and $\pm 0.013^\circ/\text{m}$ rotation without loop closure), and MoCap Ground-Truth (GT) 6D camera pose.

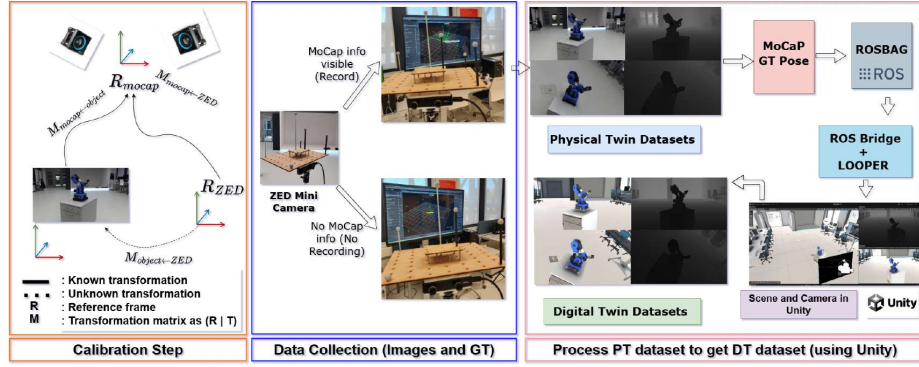


Fig. 2. Dataset generation pipeline demonstrating the workflow from PT acquisition via MoCap tracking to the generation of the paired DT sequences using the *LOOPER*.

DT Dataset Generation To create the DT counterpart, we reproduce it inside the Unity engine, providing a geometrically and visually accurate virtual environment by integrating CAD models and calibrating visual properties (colors, materials, and dimensions). By globally synchronizing the MoCap coordinate origin with Unity and leveraging a tool reference in the MoCap frame, we accurately pinpoint environmental landmarks and achieve real-to-synthetic alignment through coordinate system transformation, accounting for the handedness difference between MoCap (right-handed) and Unity (left-handed). Using LOOPER, we accurately replay all PT trajectories within the DT space. The primary advantage is isolating the visual domain shift by eliminating the trajectory gap and directly transfer human-operated real acquisition paths into virtual space with exact camera parameters (intrinsic and extrinsic). This produces strictly paired

image sequences enabling direct performance comparisons for camera localization and perfectly accurate synthetic depth maps. However, spatial misalignment between physical objects and their digital counterparts - due to measurement inaccuracies - can compromise paired data quality. We address this through an iterative correction step: virtual objects are initially placed according to the object center ground-truth, then visually adjusted by comparing the real and synthetic images until achieving satisfactory correspondence. This yields paired PT and DT images with superposition validation shown in Fig. 3.

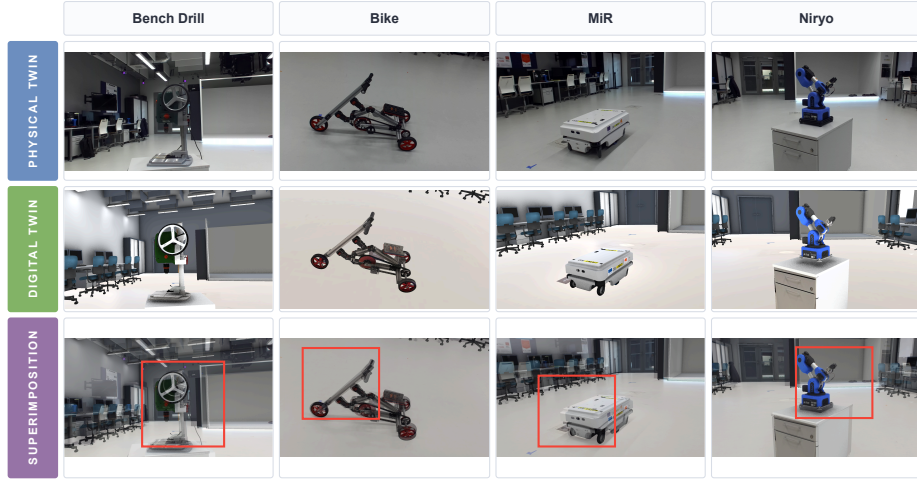


Fig. 3. Alignment comparison between Physical-Twin (PT) and Digital-Twin (DT) images. Red boxes in the superimposed view highlight visible discrepancies in alignment.

Evaluation Metrics We adopt standard evaluation metrics from recent SfM and visual localization benchmarks [10, 17, 18]. The *Registration Rate* (Reg%) reports the percentage of input images for which the pipeline estimates a 6-DoF camera pose [17, 18] and number of reconstructed 3D points (#Pts in thousands). Since SfM reconstructions live in an arbitrary coordinate frame and scale, we first align them to the MoCap GT using the closed-form Sim(3) Procrustes solution of Umeyama [19]. The *ATE* (Absolute Trajectory Error) then measures the root-mean-square Euclidean distance between aligned and Ground-Truth (GT) camera centers in meters [9]. To evaluate local pose consistency, we report *RRA* and *RTA* (Relative Rotation / Translation Accuracy), i.e. the fraction of consecutive frame pairs whose relative errors fall below $1^\circ / 5^\circ$ and $1\text{ cm} / 5\text{ cm}$ respectively [17]. We also report *mAA@5°* (mean Average Accuracy), the area under the pose-accuracy curve averaged over thresholds from 1° to 5° as defined in the Image Matching Challenge [10]. Finally, the *Runtime* corresponds to the wall-clock duration of the mapper stage only, excluding feature extraction and matching which are shared across all methods.

Environment All experiments run on a shared, high-performance compute node equipped with a single NVIDIA A100-PCIE (40 GB) GPU, an AMD EPYC

CPU, and a NAS-mounted working directory. For ease of computation, the original recorded images with 1920×1080 resolution are rescaled to 1200×675 due to memory limitations. Software dependencies are standardized within a single, reproducible Conda environment built upon Python 3.11 and the CUDA 12.9 toolchain. Core libraries include PyTorch 2.5.1 (cu121) wheels and pre-compiled COLMAP 3.13.0 and GLOMAP 1.2.0 binaries sourced directly from `conda-forge`.

5 Experimental Results

We organise the discussion below around the two research questions raised in the introduction: **RQ1**, the influence of the physical-to-digital domain gap on SfM pose accuracy and the scene properties driving it, and **RQ2**, whether the ranking of SfM pipelines is stable across industrial objects or whether the best-performing pipeline varies from one object to another. PT/DT comparisons address RQ1, and per-object cross-pipeline comparisons address RQ2.

Table 1 reports the registration rate (Reg%), the number of reconstructed 3D points (#Pts), and the mapper runtime (t) for the four SfM pipelines under sequential and exhaustive matching. COLMAP and GLOMAP consistently produce the densest reconstructions under exhaustive matching (COLMAP first and GLOMAP second on every configuration). Under sequential matching, FastMap is competitive with COLMAP, marginally densest on Bike, Niryo (both domains), and Bench Drill/DT, while COLMAP leads on Bench Drill/PT and both Mir scenes; the spread between the top two pipelines is typically only 1-4k points. *The same FastMap behaves very differently under exhaustive matching*, where it collapses to a few thousand points on Mir and on Bike/DT, even though Bike/DT was one of its strongest scenes under sequential. InstantSfM is the sparsest pipeline under sequential matching (typically 10-25% below COLMAP), while FastMap becomes the sparsest under exhaustive matching, often dropping to a small fraction of COLMAP’s point count. From a GS perspective, denser sparse clouds give a stronger initialization and reduce densification overhead, making COLMAP the safest backbone for high-fidelity reconstruction. Mapper times follow the opposite ordering. InstantSfM is the fastest mapper in five of the eight sequential-matching configurations, often substantially faster than COLMAP and several times faster than GLOMAP. FastMap is fast on PT scenes under sequential matching (notably Bench Drill/PT at 102s and Niryo/PT at 119s), but its registration rate drops to 49-62% on Mir under exhaustive matching, reflecting its sensitivity to objects that combine poor texture and symmetry, as is the case for the Mir robot. GLOMAP is consistently the slowest pipeline, with exhaustive runs reaching 1,000-1,500s on the larger datasets (Bike and Mir, 515 images each); COLMAP sits in the mid-range, while InstantSfM and FastMap are typically the fastest. For the same mapper and matcher, PT scenes generally yield more points than their DT counterparts on Bench Drill, Bike, and Niryo, with Mir as the notable exception where DT often surpasses PT. This is a first instance of the domain gap targeted by **RQ1**: the richer photometric

content of real images yields more discriminative SIFT features than the cleaner Unity renderings, so more inliers survive geometric verification.

Table 1. Registration rate (Reg%), number of reconstructed 3D points (in thousands) and mapper runtime (in seconds) for both Sequential and Exhaustive matchers across all datasets and methods. Higher ($\uparrow$) or lower ($\downarrow$) is better. Dark green = best, light green = second best (per dataset/matcher).

Dataset	Images	Method	Sequential matcher			Exhaustive matcher		
			Reg% $\uparrow$	#Pts(k) $\uparrow$	t (s) $\downarrow$	Reg% $\uparrow$	#Pts(k) $\uparrow$	t (s) $\downarrow$
BD / DT	275	COLMAP	100.0	22.8	180	100.0	31.3	344
		GLOMAP	99.6	20.7	438	99.3	17.9	687
		FastMap	100.0	24.1	246	100.0	13.3	215
		InstantSfM	100.0	18.0	110	99.6	14.9	243
BD / PT	275	COLMAP	100.0	31.6	219	100.0	36.4	440
		GLOMAP	96.7	30.6	1050	100.0	31.8	500
		FastMap	80.7	29.1	102	100.0	21.4	74
		InstantSfM	100.0	28.5	123	100.0	23.9	236
BK / DT	515	COLMAP	100.0	33.0	255	100.0	34.8	710
		GLOMAP	100.0	30.3	257	100.0	22.6	1201
		FastMap	100.0	34.2	379	95.9	7.4	478
		InstantSfM	100.0	25.5	113	100.0	15.7	534
BK / PT	515	COLMAP	100	34.5	215	100	32.4	544
		GLOMAP	99.8	32.3	471	100.2	23.3	1009
		FastMap	100.0	38.6	399	99.8	10.8	324
		InstantSfM	100.0	28.4	143	100.0	18.7	291
MR / DT	515	COLMAP	100.0	29.8	335	100.0	32.2	962
		GLOMAP	95.5	26.4	571	96.9	19.8	1427
		FastMap	100.0	12.3	819	48.9	2.1	656
		InstantSfM	100.0	22.6	767	99.6	14.1	646
MR / PT	515	COLMAP	100.0	27.6	331	100.0	26.3	646
		GLOMAP	99.6	26.3	958	100.0	20.8	1197
		FastMap	100.0	12.0	367	62.1	4.2	428
		InstantSfM	100.0	23.1	156	100.0	15.4	339
NR / DT	220	COLMAP	100	16.0	137	100	21.7	236
		GLOMAP	100.5	14.0	269	99.1	15.3	331
		FastMap	99.1	16.9	404	100.0	10.3	419
		InstantSfM	100.0	12.1	91	99.5	10.1	217
NR / PT	220	COLMAP	100.0	20.5	186	100.0	23.9	231
		GLOMAP	99.5	20.0	965	100.0	21.5	224
		FastMap	100.0	22.0	119	100.0	14.1	78
		InstantSfM	100.0	17.7	416	100.0	15.3	126

Table 2 reports ATE, RRA at $1^\circ/5^\circ$, RTA at 1 cm/5 cm, and mAA@ 5° when sequential feature matching is used. COLMAP is the only method maintaining mAA@ 5° above 85% on every scene/domain combination, with ATE leaving the centimetre range only on Mir/DT (0.36 m). GLOMAP behaves bimodally: its rotation accuracy matches COLMAP’s everywhere (RRA@5 above 97%), and on Bike/PT and Niryo/PT its ATE comes within a factor of two of COLMAP, but elsewhere it rises to 0.4 - 1.4 m because the limited-overlap sequential graph leaves the global averaging under-constrained. FastMap is occasionally close to COLMAP on Niryo/PT and Bench Drill/PT but collapses on Mir in both domains (mAA below 0.3%), showing that its sensitivity is texture-driven rather than synthetic-domain-driven. InstantSfM registers all images and recovers reasonable orientation but exhibits large absolute trajectory errors on most DT scenes, with only Niryo/DT staying competitive; its sparse, parallel design preserves local geometry yet skips the iterative global bundle adjustment needed to keep per-pair residuals from accumulating as drift. Examining

per-object rankings answers **RQ2** under sequential matching: COLMAP holds the top rank on every object/domain combination - the only pipeline that does so - while the runner-up varies (FastMap on Bench Drill/PT, GLOMAP on Bike/PT, Mir/PT, and Niryo/PT) and the worst-performer also shifts by datasets (InstantSfM on most DT datasets, FastMap on Mir in both domains). Real images contain richer texture than their Unity counterparts, producing more discriminative SIFT correspondences and substantially lower ATE on PT than DT; this texture-driven benefit is non-uniform, however: COLMAP improves 40 $\times$ in ATE from DT to PT on Mir (0.36 $\rightarrow$ 0.009 m) yet is essentially insensitive on Bench Drill.

Table 2. Pose accuracy metrics using the **sequential** matcher: ATE (m), RRA at 1 $^\circ$ and 5 $^\circ$, RTA at 1 cm and 5 cm, and mAA from 1 $^\circ$ to 5 $^\circ$. Higher ($\uparrow$) or lower ($\downarrow$) is better. Dark green = best, light green = second best.

Dataset	Method	ATE (m) $\downarrow$	RRA@1 $\uparrow$	RRA@5 $\uparrow$	RTA@1 $\uparrow$	RTA@5 $\uparrow$	mAA@5 $\uparrow$
BD / DT	COLMAP	0.0088	100.00	100.00	99.64	100.00	96.56
	GLOMAP	0.6529	98.90	98.90	10.62	89.01	60.58
	FastMap	0.6436	98.17	99.64	43.07	91.24	54.42
	InstantSfM	0.8843	96.72	98.91	5.11	94.53	13.82
BD / PT	COLMAP	0.0098	99.64	100.00	94.53	100.00	85.99
	GLOMAP	0.6011	98.49	98.87	1.13	75.09	45.66
	FastMap	0.0126	100.00	100.00	78.73	100.00	84.81
	InstantSfM	0.0748	97.08	99.64	79.56	98.54	36.52
BK / DT	COLMAP	0.0410	90.06	97.27	88.50	92.40	86.88
	GLOMAP	1.3605	89.86	97.08	0.58	10.92	72.22
	FastMap	0.6523	86.75	96.69	37.43	80.90	54.97
	InstantSfM	1.1431	77.39	95.52	0.58	55.75	17.75
BK / PT	COLMAP	0.0179	96.49	100.00	87.33	99.42	92.92
	GLOMAP	0.0254	96.09	100.00	86.69	99.41	89.42
	FastMap	0.5749	94.74	99.81	5.65	92.40	56.13
	InstantSfM	0.7775	93.76	99.61	1.36	84.02	39.12
MR / DT	COLMAP	0.3634	97.67	99.42	83.27	99.22	91.03
	GLOMAP	1.1006	93.48	98.17	3.46	88.19	59.44
	FastMap	1.5100	14.98	77.82	0.00	4.67	0.03
	InstantSfM	1.5055	87.74	93.19	0.58	6.42	7.66
MR / PT	COLMAP	0.0089	96.69	100.00	91.05	100.00	95.02
	GLOMAP	0.8147	93.56	98.24	19.53	97.27	68.05
	FastMap	1.4966	33.07	87.74	0.39	5.84	0.23
	InstantSfM	1.4055	91.25	97.28	0.58	19.65	22.95
NR / DT	COLMAP	0.0147	99.09	100.00	99.09	100.00	88.67
	GLOMAP	0.4339	98.17	99.09	44.29	97.72	85.67
	FastMap	0.4294	83.41	94.93	17.97	85.71	54.00
	InstantSfM	0.2671	91.78	99.09	58.90	95.89	17.77
NR / PT	COLMAP	0.0103	98.63	100.00	94.06	100.00	85.35
	GLOMAP	0.0143	98.62	100.00	93.58	100.00	83.45
	FastMap	0.0197	95.89	100.00	78.54	98.17	81.95
	InstantSfM	0.1987	95.43	99.09	80.82	98.17	70.72

The accuracy penalty on DT thus tracks the lack of texture - the Mir robot’s weakly textured body dominates the frame and generates weakly discriminative descriptors and fragile inlier sets - rather than a uniform “DT is harder” effect. This answers the second half of **RQ1**: image realism alone does not predict the gap; the texture of the captured object does. Under sequential matching, COLMAP is the unambiguous recommendation - the only pipeline maintaining high mAA and ATE below 0.4 m across both PT and DT domains; GLOMAP is competitive with COLMAP only on Bike/PT and Niryo/PT, where its ATE comes within a factor of two.

Table 3 re-evaluates the same four pipelines after exhaustive matching. On PT datasets, the denser correspondence graph delivers a clear improvement: COLMAP, GLOMAP, and FastMap converge to near-identical sub-centimetre ATE on Bench Drill / PT and Niryo / PT, and InstantSfM closes most of its gap on Bike / PT (mAA 39.1% seq $\rightarrow$ 89.3% exh; ATE 0.78 $\rightarrow$ 0.02 m). The most striking gain is for GLOMAP, which becomes virtually tied with COLMAP across every PT dataset (mAA $\geq$ 87%, ATE $\leq$ 0.02 m), confirming that its global formulation is correspondence-bound rather than scene-bound; the same effect even transfers to Bike / DT, where GLOMAP unexpectedly outperforms COLMAP (mAA 87.9% vs. 43.9%; ATE 0.04 m vs. 1.34 m). On the remaining DT datasets, the picture changes sharply: FastMap collapses on Mir / DT (mAA 0.14%, ATE 1.79 m) and InstantSfM also fails on Mir / DT (mAA 0.64%), while GLOMAP itself drops below its sequential baseline on Bench Drill / DT (mAA 60.6% seq vs. 34.2% exh). The cause is the lack of texture: when descriptors are weak, exhaustive matching multiplies the false-match rate and the resulting two-view geometries cannot be disambiguated by the global solvers.

Table 3. Pose accuracy metrics using the **exhaustive** matcher: ATE (m), RRA at 1° and 5°, RTA at 1 cm and 5 cm, and mAA from 1° to 5°. Higher ($\uparrow$) or lower ($\downarrow$) is better. Dark green = best, light green = second best.

Dataset	Method	ATE (m) $\downarrow$	RRA@1 $\uparrow$	RRA@5 $\uparrow$	RTA@1 $\uparrow$	RTA@5 $\uparrow$	mAA@5 $\uparrow$
BD / DT	COLMAP	0.0054	100.00	100.00	100.00	100.00	98.24
	GLOMAP	0.9151	94.48	94.85	3.31	84.93	34.18
	FastMap	0.0084	98.91	100.00	94.53	99.27	96.50
	InstantSfM	1.1697	94.50	97.07	0.73	37.00	27.81
BD / PT	COLMAP	0.0072	99.64	100.00	94.89	100.00	90.36
	GLOMAP	0.0073	99.64	100.00	94.53	100.00	89.96
	FastMap	0.0072	100.00	100.00	86.50	100.00	88.65
	InstantSfM	0.0413	96.72	98.91	87.59	98.91	57.77
BK / DT	COLMAP	1.3423	87.52	94.74	0.39	22.61	43.91
	GLOMAP	0.0409	90.06	97.27	88.69	92.40	87.90
	FastMap	0.4467	74.19	93.90	29.07	78.05	64.85
	InstantSfM	0.1329	87.14	97.47	70.17	90.64	76.54
BK / PT	COLMAP	0.0178	96.10	100.00	88.11	99.42	93.60
	GLOMAP	0.0179	95.91	100.00	86.36	99.42	93.46
	FastMap	0.6877	78.71	98.63	3.71	85.94	70.00
	InstantSfM	0.0197	93.96	100.00	73.49	98.83	89.27
MR / DT	COLMAP	1.5024	95.53	97.47	0.58	7.00	47.84
	GLOMAP	1.5051	87.95	92.37	0.60	9.04	19.97
	FastMap	1.7917	25.50	75.30	0.00	5.58	0.14
	InstantSfM	1.4919	23.63	69.92	0.59	11.72	0.64
MR / PT	COLMAP	0.0082	96.50	100.00	92.02	100.00	95.32
	GLOMAP	0.0090	96.50	100.00	90.08	100.00	94.84
	FastMap	1.6612	61.44	87.15	0.31	13.48	2.12
	InstantSfM	1.4833	78.02	89.49	0.20	11.87	20.86
NR / DT	COLMAP	0.5115	97.26	97.26	57.53	97.26	72.54
	GLOMAP	1.0334	96.76	97.22	1.39	54.63	47.94
	FastMap	1.0339	74.89	89.50	4.11	71.69	10.73
	InstantSfM	1.0260	93.12	96.79	2.29	61.93	42.71
NR / PT	COLMAP	0.0059	98.63	100.00	96.35	100.00	87.45
	GLOMAP	0.0062	98.63	100.00	94.98	100.00	87.02
	FastMap	0.0074	98.63	100.00	86.30	100.00	87.08
	InstantSfM	0.0330	94.98	100.00	85.84	96.80	71.65

Apart from Bike / DT, COLMAP remains the most consistent pipeline across the DT regime thanks to its aggressive geometric verification. Comparing Tables 2 and 3, the verdict is concrete: under exhaustive matching, COLMAP and GLOMAP con-

verge to sub-2-cm accuracy on every PT dataset, while FastMap converges only on Bench Drill/PT and Niryo/PT. COLMAP remains the recommended pipeline under both matching strategies, since it combines the highest consistency across PT and DT domains with substantially lower runtime than GLOMAP; GLOMAP matches its accuracy on PT datasets but at 1.4-2 $\times$ higher runtime, with its only unique advantage on Bike/DT, where it is the sole method to achieve sub-decimeter ATE. Mir fails not from weak texture (the physical twin reconstructs accurately) but from its large, near-symmetric body, which creates a translation/scale ambiguity where rotations are recovered while metric translations collapse. The matching strategy then decides which methods escape this ambiguity, and the digital-twin renders, lacking fine surface microtexture, degrade accuracy further in the DT domain.

Fig. 4 visualizes, for each of the four datasets, the best-performing SfM combination overlaid on the MoCap ground truth. PT and DT trajectories collapse onto the green ground truth almost perfectly, reinforcing the answer to **RQ2**: COLMAP wins on every configuration except Bike/DT, where GLOMAP is the best performer. Note that on Mir/DT and Niryo/DT, sequential matching is the winning configuration because the symmetric and weakly textured DT surfaces produce ambiguous correspondences under exhaustive matching.

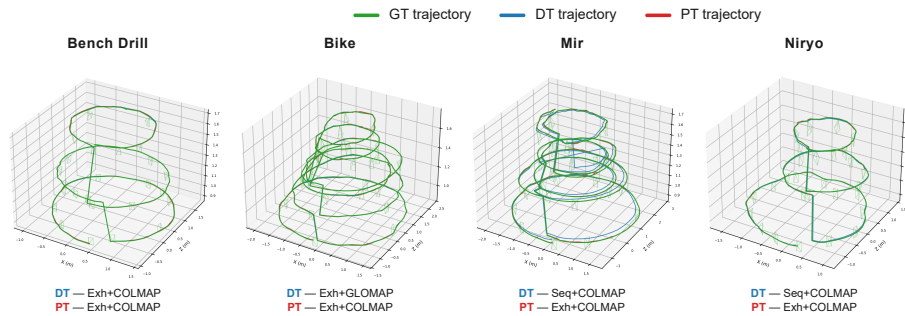


Fig. 4. 3D trajectory comparison for the four industrial datasets showing the best-performing SfM combination overlaid on the MoCap GT (green). The matcher and SfM method selected for each domain are listed under every plot.

Fig. 5 plots mapper runtime against the number of reconstructed 3D points, with separate panels for sequential (left) and exhaustive (right) matching.

The clearest pattern is a separation between COLMAP and GLOMAP, which produce denser reconstructions at higher runtimes, and FastMap and InstantSfM, which produce fewer points at lower runtimes. Switching from sequential to exhaustive raises the runtime of COLMAP, GLOMAP, and InstantSfM but rarely adds proportionally more points; FastMap is the exception, showing no consistent speed-up under exhaustive and a sharp drop in point count, most notably on Mir. The per-method regression lines (solid for DT, dashed for PT) make the physical-to-digital gap visible per pipeline: the wider the separation between solid and dashed lines, the larger the gap for that method. Around 10-15k well-distributed points is typically enough for clean Gaussian Splatting initialization. COLMAP stays comfortably above this threshold on every configuration, GLOMAP on most, FastMap on the non-Mir sequential configurations, and InstantSfM hovers near the threshold on the smallest datasets.

Taken together, the experiments answer both research questions. **RQ1**: The physical-to-digital domain gap is real but highly non-uniform - severe on Mir, modest on Bike

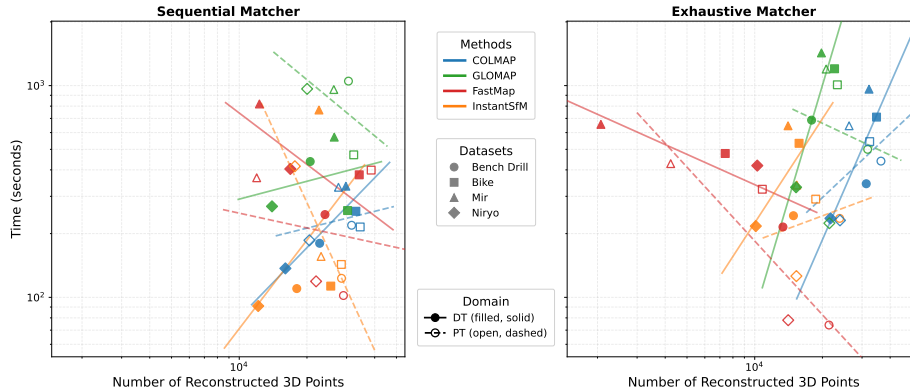


Fig. 5. Mapper runtime vs. reconstructed 3D points for four SfM pipelines under sequential (left) and exhaustive (right) matching.

and Niryo, essentially absent on Bench Drill. The main driver is not visual realism but *the texture of the captured object*: the Mir robot’s weakly textured body fills most of the frame and yields weak SIFT descriptors and fragile inlier sets. **RQ2:** The best-performing pipeline is stable across all four objects - COLMAP wins on every configuration under sequential matching and on seven of eight under exhaustive matching (the lone exception is Bike/DT, where GLOMAP wins). The ranking of the remaining three pipelines varies sharply by object: GLOMAP excels on textured PT datasets but degrades on DT; FastMap works on Bench Drill and Niryo/PT, is mid-range on Bike, and collapses on Mir; InstantSfM drifts on Mir in both domains but recovers strongly on PT scenes under exhaustive matching. COLMAP is therefore the recommendation emerging from both questions: the only pipeline whose top-rank position is object-independent.

6 Conclusion and Perspectives

We introduced a paired Physical-Twin (PT) and Digital-Twin (DT) industrial dataset with Motion-Capture (MoCap) Ground Truth (GT) and used it to benchmark four SfM pipelines (COLMAP, GLOMAP, FastMap, InstantSfM) under sequential and exhaustive matching across four industrial objects. The evaluation reports registration rate, sparse-point count, runtime, ATE, RRA, RTA, and mAA, providing a reproducible reference for camera localization. The main findings are: (i) COLMAP remains the most accurate and reliable SfM backbone for both PT and DT data; (ii) GLOMAP closely matches COLMAP on most PT datasets under exhaustive matching, but generally requires higher mapper runtime and is more sensitive to the correspondence graph; (iii) InstantSfM consistently reduces mapper runtime relative to COLMAP and FastMap becomes faster under exhaustive matching, but the pose accuracy of both pipelines is unreliable on the poorly textured and symmetric DT objects; (iv) Sequential matching is safer for DT datasets with weak texture or symmetry and exhaustive matching is more beneficial on PT acquisitions and on textured DT scenes. Exhaustive matching helps most pipelines but its effect is method-dependent; and (v) the domain gap is measurable but not uniform across objects. The gap is largest on the weakly

textured and symmetric objects indicating that texture richness and object geometry are stronger drivers than the real-versus-synthetic domain label alone.

Perspectives. Three complementary research directions follow from this study, all centred on advancing 3D reconstruction and 6D camera pose estimation models for industrial scenes. (i) The proposed benchmark can serve as input to radiance field rendering pipelines such as 3DGS: extending the protocol to representative GS variants would help characterize the extent to which the SfM front-end limits photorealistic reconstruction in terms of fidelity-per-second, and support a more detailed analysis of how the physical-to-digital domain gap propagates. (ii) The paired PT/DT trajectories make this dataset well suited to investigate hybrid PT + DT training strategies, in which digital-twin sequences augment real acquisitions; evaluating the resulting hybrid models on physical-twin data would help assess how much synthetic supplementation can reduce the texture-driven accuracy gap observed on weakly textured industrial objects such as Mir. (iii) Building on this hybrid setting, the DT pipeline could enable a “train on DT, deploy on PT” workflow that may substitute MoCap-based acquisitions with Unity-replayed sequences, supporting more scalable industrial deployment of 3D reconstruction and pose estimation pipelines while preserving geometric fidelity.

Acknowledgements This work is being carried out as part of the FUSION (Framework for Universal Software Integration in Open robotics) project, funded through the "i-Démo régionalisé" call for projects launched by the French government and the Normandy region, as part of the France 2030 program. This work is partially supported by PHC BALATON REMOTED (Robotic EnvironMent hybridizatiOn rendering based-on data-efficient radiance field TEchnologies and Digital twin) 52934RC project / TÉT project No. 2024-1.2.5-TÉT-2024-00019, supported by the National Research, Development and Innovation Office of Hungary.

References

1. Qiang Qin, Zhihao Liu, Ruirui Zhong, Xi Vincent Wang, Lihui Wang, Magnus Wiktorsson, and Wei Wang. Robot digital twin systems in manufacturing: Technologies, applications, trends and challenges. *Robotics and Computer-Integrated Manufacturing*, 97:103103, 2026. doi:10.1016/j.rcim.2025.103103.
2. Meng Xu, Youchen Wang, Bin Xu, Jun Zhang, Jian Ren, Zhao Huang, Stefan Poslad, and Pengfei Xu. A critical analysis of image-based camera pose estimation techniques. *Neurocomputing*, 570:127125, 2024. doi:10.1016/j.neucom.2023.127125.
3. Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: A comprehensive survey. *International Journal of Computer Vision*, 2026. doi:10.1007/s11263-025-02646-6.
4. Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. doi:10.1145/3592433.
5. Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16739–16752, June 2025. doi:10.1109/CVPR52734.2025.01560.

6. Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), July 2017. doi:10.1145/3072959.3073599.
7. Steve Borkman et al. Unity perception: Generate synthetic data for computer vision. *arXiv*, 2021. doi:10.48550/arXiv.2107.04259.
8. Paul Bréon, Imane Argui, and Vincent Havard. LOOPER: A framework for synthetic dataset generation with configurable sensors and multi-view XR environments. In *Extended Reality: International Conference, XR Salento 2026*, page 20, Otranto, Italy, June 2026. Springer-Verlag.
9. Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, 2017. doi:10.1109/CVPR.2017.272.
10. Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiří Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. doi:10.1007/s11263-020-01385-0.
11. Tomas Hodan et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–35, 2018. doi:10.1007/978-3-030-01249-6_2.
12. Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888, 2017. doi:10.1109/WACV.2017.103.
13. Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. doi:10.48550/arXiv.1711.00199.
14. Sunil Choudhary, Nicolas Ragot, Vincent Havard, and Yohan Dupuis. Evaluating Robustness of 3D Gaussian Splatting-Based 6D Camera Pose Refinement Under Degraded Conditions for Lightly Textured Industrial Synthetic Objects. In *IECON 2025 – 51st Annual Conference of the IEEE Industrial Electronics Society*, pages 1–7, Madrid, Spain, October 2025. IEEE. doi:10.1109/IECON58223.2025.11221686.
15. Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, June 2016. doi:10.1109/CVPR.2016.445.
16. Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 58–77, 2024. doi:10.1007/978-3-031-73661-2_4.
17. Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. Fastmap: Revisiting dense and scalable structure from motion. 2025. doi:10.48550/arXiv.2505.04612.
18. Jiankun Zhong, Zitong Zhan, Quankai Gao, Ziyu Chen, Haozhe Lou, Jiageng Mao, Ulrich Neumann, Chen Wang, and Yue Wang. InstantSfM: Towards gpu-native sfm for the deep learning era. 2025. doi:10.48550/arXiv.2510.13310.
19. Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi:10.1109/34.88573.