iMED: A Multi-Endoscope Dataset for Surgical 3D Perception

Sierra Bonilla[1], Fengyi Jiang[2], Chinedu Nwoye[2], Jingpei Lu[2], Kailey Reardon[2], Humphrey W. Chow[2], Francisco Vasconcelos[1], Sophia Bano[1], Adam Schmidt[2], and Omid Mohareri[2]

[1] University College London, London WC1E 6BT, United Kingdom
[2] Advanced Research, Intuitive, Sunnyvale, CA 94086, USA
sierra.bonilla.21@ucl.ac.uk

Abstract. We introduce iMED, the first synchronized multi-endoscope dataset for robot-assisted minimally invasive surgery (RAMIS), designed to address a fundamental limitation in surgical vision benchmarking: the absence of independent held-out viewpoints. Existing surgical datasets capture single-trajectory sequences from a single endoscope, making it impossible to distinguish geometric generalization from photometric interpolation along a narrow forward-facing path. iMED provides 340 sequences (≈170K synchronized timepoints with 4 views per timepoint) recorded simultaneously from two independent stereo endoscopes across ex vivo, postmortem, and live surgical settings, spanning 14 specimens with diverse anatomical regions and motion regimes. The dataset includes calibrated camera intrinsics, ArUco-based frame-wise pose estimates with uncertainty quantification, instrument segmentation masks, and rich clinical metadata. Using a train-on-one-endoscope, test-on-another evaluation protocol, we benchmark 23 state-of-the-art methods across rigid and deformable novel view synthesis, pose estimation, feature matching, and monocular depth estimation. Our experiments show that methods relying primarily on photometric supervision degrade substantially under this held-out-endoscope setting, while methods with explicit geometric regularization are more robust across views in our benchmark. Notably, large-scale foundation models transfer surprisingly well to surgical imagery, often outperforming domain-specific models, suggesting pitfalls in current fine-tuning protocols. iMED establishes a new evaluation protocol for geometric generalization in surgical vision. Dataloader and dataset links can be found at `github.com/surgical-vision/imed`.

Keywords: Surgical vision · Robot-assisted minimally invasive surgery · 3D from multi-view

1 Introduction

Non-rigid scenes with constrained camera motion are among the most challenging regimes for image-based 3D perception tasks (i.e. pose estimation, feature matching, optical flow, depth estimation, and novel view synthesis (NVS)).

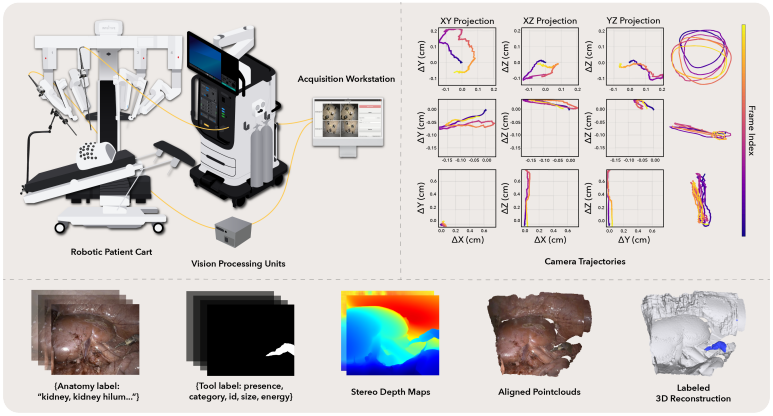

Fig. 1: Data platform with two stereo endoscopes (top-left), common camera trajectories (top-right), and pipeline outputs including session/anatomy/instrument/motion labels, camera intrinsics, ArUco pose estimates, inferred tool masks , stereo depth (code provided), and derived 3D geometry. See Tab. 2 for provenance.

Although challenging, such conditions often arise in real-world robotic applications like digital human modeling of clothed bodies [7], submarine exploration [10], and robot-assisted minimally invasive surgery (RAMIS) [53]. Among these, RAMIS represents an extreme regime. Cameras operate inside the body under severe viewpoint constraints, typically following narrow, forward-facing trajectories through deformable tissue with strong specularities and transient phenomena (smoke, bleeding, topology changes). These conditions compress multiple failure modes for 3D perception methods, making RAMIS an informative stress test for visual geometry in real-world robotic systems.

Computer vision benchmarks have historically advanced 3D perception through standardized evaluation protocols [20]. In rigid environments, datasets capturing scenes from many surrounding viewpoints with known camera poses enabled progress in multi-view reconstruction and mapping [1,12,30,74]. Dynamic-scene benchmarks later extended these ideas to time-varying NVS and reconstruction [27,44,45,50], typically relying on wide-baseline or 360°/180° capture where cameras observe scenes from diverse directions. Surgical vision datasets have enabled progress in reconstruction, camera tracking, and tool analysis [3,22,40,53]. Unlike the surrounding-view capture in general benchmarks, endoscopic cameras follow narrow forward-facing paths through tissue, producing sequences with significant image overlap and limited viewpoint diversity.

In NVS and 3D scene representation, models are typically evaluated using images at held-out camera poses and reporting metrics such as PSNR and SSIM against held-out views [5, 23, 29, 55, 72]. These metrics quantify photometric fidelity and have become standard quantitative benchmarks. However, they are proxies for reconstruction quality rather than direct measures of geometric consistency: high PSNR/SSIM can reflect local interpolation between nearby view-

Table 1: Comparison of stereo-endoscopic datasets for 3D perception in minimally invasive surgery. Multi-Camera* defined as more than one stereo camera pair. **only for 2 phantom validation sequences with CT scans.

Dataset	Dataset Attributes					Scene Types				Inferred Labels		Provided Reference Signals	
	Seq.	Specimens	Modality	Setting	Live	Static	Dyn.	Rigid	Deform.	Obj.	3D Model	Tool Masks	Camera Calib.	Multi-Camera*
EndoVis 17/18 [2, 4]	29	29	Da Vinci Xi Da Vinci X	*In vivo*	✓		✓		✓			✓		
MITI [22]	-	1	Karl Storz 3D Tipcam	*In vivo*	✓				✓				✓	
STIR [53]	572	4	Da Vinci Xi	*Ex vivo* *In vivo*	✓				✓				✓	
Hamlyn Centre [42, 49, 58]	-	-	Da Vinci Si	Phantom *Ex vivo* *In vivo*	✓	✓			✓		✓**			
SERV-CT [16]	16	2	Da Vinci	Postmortem		✓					✓		✓	
SCARED [3]	9	9	Da Vinci Xi	Postmortem		✓					✓		✓	
StereoMIS [40]	16	6	Da Vinci Xi	*In vivo*	✓		✓		✓			✓	✓	
iMED (Ours)	340	14	Da Vinci 5	*Ex vivo* *In vivo* Postmortem	✓	✓	✓		✓		✓	✓	✓	✓

points, without uniquely determining whether the model has learned the underlying 3D structure [67]. This limitation is amplified when train and test poses lie along the same forward-facing trajectory, where test views are typically sampled from nearby frames in time so differ only slightly in view. A complementary evaluation is to test generalization under larger viewpoint shifts. In classic multi-camera dynamic-scene benchmarks, this protocol is well established: models are trained on a subset of cameras and evaluated on held-out cameras [8, 27, 33].

To address this limitation in surgical vision, we introduce **iMED**, the first multi-endoscope dataset for RAMIS, providing synchronized recordings from two physically independent stereo endoscopes, four camera streams with overlapping fields of view, enabling cross-trajectory geometric validation (see Fig. 1). The dataset comprises 340 sequences totaling ∼170K synchronized timepoints (4 frames×60 FPS) across diverse surgical scenarios: 10 *ex vivo* organ specimens (chicken, porcine, and bovine; ≈7 sequences each), 3 human cadaver subjects (≈6 anatomical sites per subject, ≈7 sequences per site), and 1 live porcine subject (10 anatomical sites, ≈7 sequences per site). Data collection was performed in a USDA-licensed clinical lab using one porcine model and three human cadaver models sourced from a willed body program. Each 10-15 second sequence captures varied conditions including static scenes, respiratory motion, tool-tissue interactions with 20 common robotic instruments (see both Tab. 2 and Suppl.Tab.7, contact-based tissue deformations, lighting variations, and specularities. All sequences include comprehensive metadata: anatomical regions labeled by clinical specialists with surgical robotics expertise, robot tool types, powered instrument usage (e.g., electrocautery), and motion categories (e.g., camera sweep, robot tool-tissue interaction, calibration). All frames include inferred filtered instrument segmentation masks generated by [9]. Refer to Tab. 2 for label provenance. To filter sequences towards specific applications (e.g. deformable vs. rigid), refer to Suppl.Tab.10. Recordings include intrinsic calibrations and frame-wise ArUco-based [18] camera poses refined through reprojection-error optimization with uncertainty estimates.

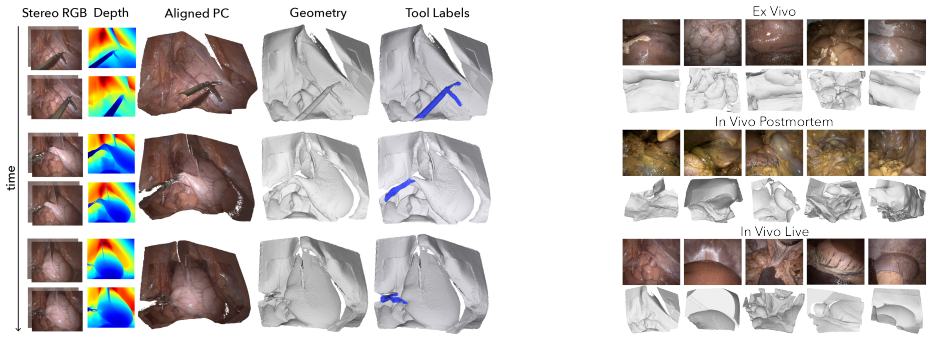

Fig. 2: Example geometric reconstructions and label projection enabled by iMED. *Left:* synchronized stereo views from both endoscopes, stereo depth estimationo [66], aligned point clouds from provided camera calibrations, Poisson-reconstruction, and reprojected tool masks. *Right:* Representative anatomical regions with surface reconstructions from *ex vivo* specimens (chicken, porcine, bovine), human cadaver sites, and live porcine tissue. Inference code to recreate geometry and 3D tool labels is provided.

Using iMED, we evaluate 23 SOTA algorithms across tasks central to 3D perception: rigid NVS, deformable (online & not) NVS, camera pose estimation, feature matching, and depth estimation. Our analysis suggests that, under this held-out-endoscope protocol, methods relying primarily on photometric supervision degrade under physically independent viewpoints, while methods with explicit geometric regularization are more robust across views. Depth supervision has representation-dependent effects, and large multi-domain foundation models often transfer well to surgical imagery compared with surgical-only training or aggressive fine-tuning. We define train-on-one-endoscope, test-on-another evaluation protocols that eliminate trajectory overlap and measure geometric generalization across physically independent cameras, a standard in multi-view human capture datasets that is absent from surgical benchmarks. Motivated by the lack of independent test views in surgical datasets, we introduce iMED and make the following contributions:

1. A train-on-one-endoscope, test-on-another evaluation protocol providing held-out viewpoints with large baseline and viewing-direction differences, as a better proxy for geometric generalization beyond photometric interpolation
2. A synchronized multi-endoscope dataset for RAMIS with 340 sequences (≈170K timepoints with 4 images per timepoint) across *ex vivo, in vivo* postmortem, and *in vivo* settings (See Fig. 3)
3. Calibrated multi-view images and structured annotations (cam intrinsics, reference trajectories, anatomical labels, motion categories, instrument masks)
4. A benchmark of 23 methods across NVS, pose estimation, feature matching, and monocular depth estimation that exposes previously unobserved behaviors: models trained using only image supervision perform poorly on new viewpoints, methods with explicit 3D regularization generalize more reliably, and foundation models often transfer better than surgical-only training

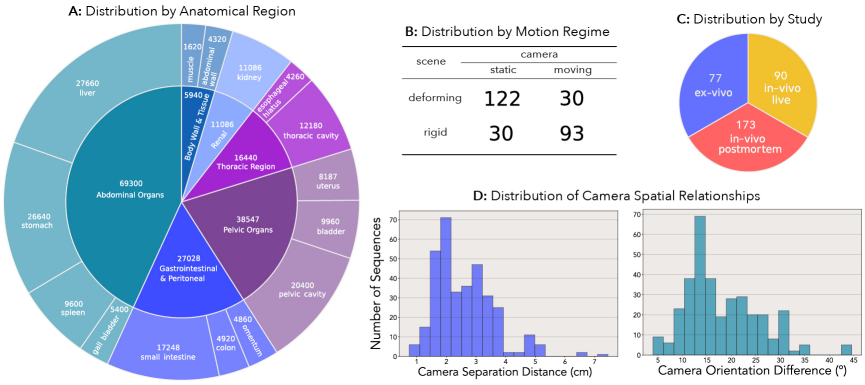

Fig. 3: *A:* Distribution of seqs. by anatomical region and organ type including number of timepoints per category. *B:* Distribution of seqs. by motion regime (camera motion vs. scene motion). Calibration-only sequences are excluded. *C:* Distribution of seqs. by session type. *D:* Distribution of seqs. by cam spatial relationships, showing inter-cam separation distance and orientation difference.

2 Related Work

Multi-View and 3D Reconstruction Datasets. Rigorous evaluation of 3D reconstruction requires validated geometric ground truth. Multi-view stereo benchmarks establish ground truth through controlled conditions: DTU [1] employs robotic arms capturing 49-64 viewpoints with trajectories; BlendedMVS [73] synthesizes depth maps from 20-1000 input views per scene via photogrammetric reconstruction; Tanks and Temples [30] validates outdoor reconstructions from 150-300+ images against laser scans. These methods enable pixel-accurate depth evaluation but are confined to static, well-textured scenes. Object-centric datasets (CO3D, uCO3D [37, 52]) capture objects from smartphone video, while unbounded scene datasets like Mip-NeRF 360 [5] and RealEstate10K [76] use COLMAP [54] and ORB-SLAM2 [43] on controlled captures. In many of these datasets, static scenes are recorded from a wide range of viewpoints, often circling around the object or environment. Because the same scene is observed from many directions, it is logical to train models on some views and evaluate them on other viewpoints from the same capture.

Dynamic scenes introduce an additional challenge: views captured at different moments no longer correspond to the same geometry. Several datasets study deformable reconstruction under these conditions. D-NeRF [50] and Nerfies [44] explore neural rendering from monocular video, while HyperNeRF [45] uses a stereo rig. These works typically focus on predicting intermediate time steps, training on subsets of frames and testing on unseen frames along the same trajectory. Multi-view human capture datasets instead measure spatial generalization directly by recording scenes with many synchronized cameras. Panoptic-Studio [27, 28] (480 cameras), Immersive Video Dataset [8] (46 cameras), and Plenoptic Video [33] (21 cameras) train models on some viewpoints and evalu-

Table 2: Label provenance for iMED. We report annotation level, generation type, source method, and whether the annotation is included in the public data release. For storage reasons, stereo depth and 3D geometry are reproducible via provided code.

Label	Level	Provenance	Source / Method	Released
Session type	Seq.	Manual	Curator	✓
Broad anatomical region	Seq.	Manual	Clinical specialist	✓
Detailed scene caption	Scene	Manual	Clinical specialist	✓
Instrument type	Seq.	Manual	Clinical specialist	✓
Motion regime/movement type	Seq.	Manual	Curator + protocol	✓
Camera intrinsics/stereo baseline	Seq.	Measured	Calibration procedure	✓
Frame-wise camera poses	Frame	Measured	ArUco PnP + refinement	✓
Instrument masks	Frame	Inferred	[9]	✓
ArUco masks	Frame	Inferred	[9]	✓
ArUco in-painted	Frame	Inferred	[34]	✓
Stereo depth	Frame	Inferred	[66] (code provided)	×
3D geometry (point clouds/meshes)	Frame	Derived	Poisson (code provided)	×

ate them on others captured at the same moment in time. Because the held-out views come from different cameras, this setup provides a stronger test of the geometry than interpolation along a single trajectory. These datasets assume isolated deforming subjects against largely static backgrounds, a configuration incompatible with surgical scenes where the entire field of view may deform. The Drunkard's Dataset [51] further studies deformable visual odometry under exploratory camera motion using synthetic data designed with endoscopic navigation in mind, where the entire scene undergoes non-rigid distortion.

Surgical Datasets. Surgical datasets face unique ground truth challenges. The physical impossibility of placing dozens of cameras inside the body constrains surgical vision systems to far fewer synchronized viewpoints than multi-view benchmarks. Early surgical datasets established ground truth through external sensing modalities: SCARED [3] uses structured light across 9 sequences; SERV-CT [16] aligns endoscopic views with CT scans but provides only 16 static image pairs from two specimens; EndoAbS [46] employs laser scanning for 120 image pairs across 20 fixed phantom poses. These datasets, captured from postmortem tissue or phantoms, enable stereo depth evaluation but are fundamentally limited to single-trajectory, single-timepoint capture in non-deforming environments. Recent datasets expanded temporal and deformation coverage while maintaining single-trajectory capture: STIR [53] provides 572 sequences with infrared marker tracking; StereoMIS [40] contributes 16 sequences with poses; MITI [22] features stereo with calibrations within one procedure; SegSTRONG-C [14] targets tool segmentation providing binary instrument masks over a small set of sequences (17) *ex vivo* animal tissues. Despite increased sequence counts, all existing surgical datasets provide single endoscope captures along a single path through time. iMED fills these gaps (real-world deformation, multi-camera capture) for surgical settings through synchronized dual-endoscope capture (4 camera streams) with overlapping fields of view, enabling cross-trajectory spatial validation across 340 sequences spanning *ex vivo*, postmortem, and live surgical scenes (examples of anatomical regions in Fig. 2) with ArUco-based reference pose estimation.

Fig. 4: *Left:* ArUco calibration cube design iterations leading to our final 8 mm wooden cube design. *Right:* ArUco-based pose estimation pipeline: (1) marker detection, (2) PnP pose estimation, (3) relative transformation via RANSAC and refinement.

3 Dataset Curation

3.1 Acquisition Set-up

Data was acquired using three Anonymized surgical systems with three 8mm 0° and four 8mm 30° endoscopes. Due to hardware constraints, only one endoscope mounts to the robotic arm operationally. For *in vivo* sessions, the second endoscope was placed on an articulated passive arm (See Fig. 1, top left); for *ex vivo* sessions, both endoscopes were hand-guided in non-operative configuration. Throughout this paper, we designate the stationary endoscope mounted on the articulated arm as *Endoscope 1*, which remains fixed during data collection. The mobile endoscope mounted on the robotic patient cart with active illumination is designated as *Endoscope 2*. Only *Endoscope 2*'s integrated light source was activated during acquisition to maintain lighting conditions representative of standard MIS procedures. Since the two endoscopes observe different illumination characteristics, we normalize their photometric appearance using sequence-level statistics to ensure robustness against per-frame artifacts. See Suppl. Sec. 6.4 for evaluation of multiple normalization strategies.

We capture four HDMI streams from two hardware-synchronized endoscopes separated by 2.6±1.2 cm with 18.7°±10.0° orientation difference (Fig. 3). This 2.6cm inter-camera separation is substantially larger than the 4mm intra-endoscope stereo baseline, representing up to 50% of the typical 6-13 cm laparoscopic field of view [60,64]. Synchronization accuracy across ≈170K frame pairs achieved mean timing difference of 3.36 ms (SD = 3.90 ms), much less than the ≈17ms frame duration. See Suppl. Sec. 6.1 for complete hardware synchronization methodology and timing analysis. All sequences are compressed using H.264 CRF 18 encoding; we recommend GPU-accelerated decoding (NVDEC) for efficient processing.

Calibration Strategy Due to hardware constraints preventing direct capture of relative pose between the two endoscopes, we developed an external calibration system using custom calibration cubes with ArUco fiducial markers [18]. ArUco markers are square fiducial tags with an internal bit pattern designed for pose estimation and unique identification. Each marker consists of an outer perimeter of black cells surrounding an inner data payload region (examples seen in Fig. 4). Our design requirements included: (1) fit through standard trocar ports (≤12mm [35]), (2) provide high-contrast markers for reliable detection,

(3) be manipulable by robotic instruments, (4) remain stable via surface tension when placed on tissue, and (5) resist staining during acquisition sessions. After multiple design iterations (Fig. 4), we converged on 8mm wooden cubes with 5.7mm laser-etched ArUco markers from the `"ARUCO_MIP_36h12"` dictionary [19]. See Suppl. Sec. 6.1 for complete design iteration analysis.

We use two acquisition protocols, filterable via metadata flags: (1) *Static Camera Post-Calibration* (≈40% of sequences): ArUco cubes are moved around to densely sample the working volume, then removed before tissue manipulation sequences proceeds with fixed cameras (mean drift pre and post tissue manipulation: 1.97° rot, 3.3mm trans, indicating stable calibration throughout manipulation tasks); (2) *Dynamic Camera Sequences* (≈40% of seqs): markers remain visible throughout to enable continuous pose estimation during movement.

3.2 Reference Camera Poses

To provide the reference camera trajectories, we employ an ArUco marker-based pose estimation pipeline computing relative transformations between endoscopes (Fig. 4). For each 5.7mm ArUco marker detected in both cameras, we estimate 6-DOF poses using a Perspective-n-Point solver [26] with known marker geometry and camera intrinsics. We aggregate observations across a sliding window and compute rel. camera-to-camera transformations for commonly visible markers.

The full pipeline consists of five stages: (1) *RANSAC Initialization and Refinement:* To handle detection errors and occlusions, we use RANSAC [17] to identify inlier observations based on angular and translational consistency thresholds. We then refine the best hypothesis by minimizing bidirectional reprojection error across all inlier marker corners using Levenberg-Marquardt optimization [32, 39]. (2) *Stereo Fusion:* Each endoscope contains left and right stereo cameras. We independently compute left-to-left and right-to-right transformations, then average via quaternion interpolation for the final trajectory. For static calibration sequences, the sliding window spans the entire sequence duration, providing a globally-optimized transformation from all observations. (3) *Outlier Detection:* We identify outliers through frame-to-frame velocity analysis. Expert laparoscopic tool velocities are less than 8.5 cm/s [25]. At 60 fps capture rate, this corresponds to frame-to-frame displacements of approximately 0.14 cm. Considering this and a working volume of 20cmx20cmx20cm, we set conservative thresholds rejecting poses with translation changes >0.7 cm/frame and rotation changes >3.0°/frame to effectively filter ArUco detection errors and mechanical artifacts while preserving plausible motion. (4) *Temporal Smoothing:* We apply Savitzky-Golay filtering ($W = 15$ frames, $p = 1$) to translation axes and smooth rotations in quaternion space with sign continuity enforcement and renormalization. Outlier poses use linear interpolation (translation) and SLERP [57] (rotation) between smoothed neighbors. (5) *Uncertainty Quantification:* We assess pose reliability with consistency between independently computed left-to-left and right-to-right transformations, quantifying translational (in m) and rotational (in °) uncertainty. Ablation studies validate each stage of the trajectory pipeline, showing up to 90% reprojection err. reduction and sub-2 mm physical

Table 3: Reference trajectory accuracy after full refinement across session types. See Supplementary Sec. 6.5 for complete ablations and 3D consistency metrics.

Session Type	Reproj. Err. (px)↓	Physical Err. @5cm (mm)↓
Ex vivo	38.33	1.88
In vivo (postmortem)	17.00	0.83
In vivo (live)	11.87	0.58

accuracy at 5 cm working distance. We release trajectories, raw calibration seqs, and recomp code; summary results appear in Tab. 3, (ablations in Suppl 6.5).

4 Benchmarking Experiments

We evaluated 23 SOTA methods: two rigid NVS, six deformable NVS, 11 feature matching/pose estimation methods, and four monocular depth estimation methods. Perhaps not surprisingly, methods designed for wide baseline multi-view datasets struggle when confronted with surgical constraints. The train-on-endoscope-2, test-on-endoscope-1 protocol exposes a failure mode: methods that interpolate photometric appearance collapse when forced to extrapolate geometric structure from constrained viewpoints. Performance correlates with geometric regularization, but at the cost of computation time, a challenge for real-time constrained 3D perception. All experiments were run on four NVIDIA RTX 6000 Ada GPUs. For reproducibility, the sequences are labeled (Suppl. Tab. 10). For the NVS, pose, and feature-matching tasks, reference cam-to-cam transforms are estimated from calibration observations in the static-camera protocol; dynamic-camera trajectories are released as a resource and can be filtered using metadata.

4.1 Rigid NVS

Methods. NeRF [41] and 3D Gaussian Splatting (3DGS) [29] represent current SOTA for rigid NVS: NeRF employs implicit volumetric rendering through coordinate-based MLPs, while 3DGS uses explicit anisotropic Gaussians with differentiable rasterization. We evaluate with and without depth supervision from [66, 70] to test if geometric priors from surgical imagery improve reconstruction quality. All experiments were conducted on processed [54] test sequences.
Results. Depth supervision produces opposite trends across representations: NeRF degrades, while 3DGS improves (Tab. 4). For NeRF, depth supervision leads to ghost-like, semi-transparent artifacts and lower fidelity at the test viewpoint (Fig. 5). To analyze this behavior, we interpolate from a train camera toward a held-out test camera in 10% increments and compute SSIM between the rendered images and the test held-out view. In the expected case, SSIM should increase monotonically as the rendering approaches the test camera pose, which is observed for vanilla NeRF. However, dep-sup NeRF exhibits non-monotonic SSIM trends, where geometric convergence does not yield better photometric values.

Table 4: NVS benchmarks. Methods trained on *Endoscope 2* and evaluated on *Endoscope 1*. Rigid and deformable methods perform offline optimization vs. online methods incrementally update the scene using only the most recent observations during acquisition (see Sec. 4.1). FPS excludes SfM [54] preprocessing.

Method	Depth	PSNR↑	SSIM↑	Render (FPS)↑	Train (FPS)↑
Rigid Novel View Synthesis					
NeRF [41]	x	17.24	0.65	0.11	0.21
NeRF [41]	DA [70]	16.29	0.55	0.11	0.19
NeRF [41]	FS [66]	15.93	0.55	0.11	0.18
3DGS [29]	x	14.82	0.59	487.62	0.59
3DGS [29]	DA [70]	14.77	0.63	509.99	0.36
3DGS [29]	FS [66]	15.33	0.63	497.43	0.31
Deformable Novel View Synthesis					
4DGS [71]	x	13.74	0.57	578.46	0.38
D-3DGS [72]	x	14.50	0.51	31.41	0.14
4DGS [68]	x	15.65	0.33	38.99	0.76
Endo-4DGS [24]	DA [70]	16.26	0.59	309.37	0.66
Online Deformable Novel View Synthesis					
Hayoz [23]	RAFT [61]	15.66	0.40	224.15	0.15
Hayoz [23]	FS [66]	15.92	0.40	229.39	0.14
DynOmo [55]	DA [70]	24.16	0.87	65.18	0.07
DynOmo [55]	FS [66]	25.29	0.90	63.58	0.06

Prior work [62] suggests NeRF may "cheat" by modeling semi-transparent volumes to approx. reflections; dep. sup. may accentuate this behavior, producing artifacts that worsen as views diverge from the train distribution.

4.2 Deformable NVS

Methods. Live surgical scenes exhibit full field-of-view non-rigid deformation from respiratory motion, cardiac pulsation, peristalsis, and tool-tissue contact. We benchmark six deformable Gaussian Splatting variants. *Offline methods* optimize over complete sequences: D-3DGS [72], 4DGS (Wu et al.) [68], 4DGS (Yang et al.) [71], and Endo-4DGS [24]. *Online methods* perform incremental optimization during acquisition: Hayoz et al. [23] and DynOmo [55]. We evaluate Hayoz with both its original RAFT [61] depth prior and FoundationStereo [66]; DynOmo with both its original Depth Anything [70] and FoundationStereo.

Results. Performance correlates directly with geometric regularization strength (Tab. 4, Fig. 6). Methods relying solely on photometric supervision fail under scenarios with constrained viewpoints, where photometric cues alone cannot disambiguate geometric structure. D-3DGS achieves low performance using only L1+D-SSIM photometric loss, allowing unconstrained per-Gaussian warping that produces geometrically inconsistent reconstructions across independent viewpoints. The 4DGS variants add temporal regularization through total variation constraints on hexplane/temporal features but lack explicit spatial coherence, achieving marginal improvements. Methods incorporating explicit 3D

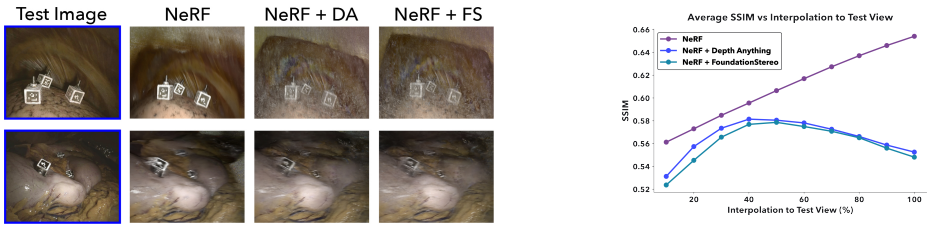

Fig. 5: *Left:* Rendered test views show ghost-like artifacts in depth-supervised variants. *Right:* SSIM computed against the test held-out image during camera interpolation from a training (0%) to test (100%) viewpoint. NeRF (purple) improves monotonically, while depth-supervised variants degrade despite geometric convergence, suggesting transparent density surfaces rather than opaque geometry.

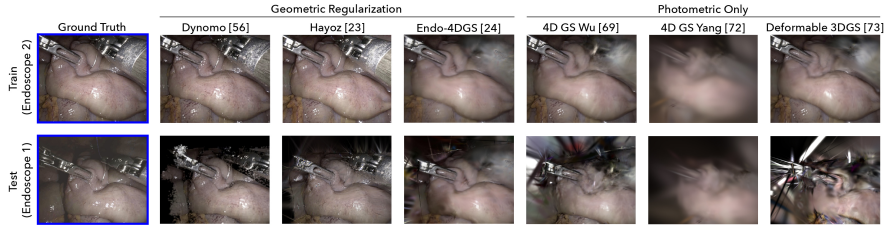

Fig. 6: Qualitative comparison of the Deformable NVS methods.

geometric constraints demonstrate substantial gains. Endo-4DGS achieves better quality through depth supervision, surface normal consistency, and confidence weighting—directly enforcing geometric plausibility. Hayoz applies local rigidity to sparse anchor points, providing modest spatial coherence. DynOmo achieves superior performance, applying dense 3D constraints to all Gaussians: local rigidity in position and rotation, isometry for long-term distance preservation, feature-similarity-based neighbor selection, and instance-aware grouping.

The train-on-one, test-on-another protocol measures geometric fidelity instead of photometric overfitting. Since methods without 3D constraints exhibit quality collapse at test views (Fig. 6), this suggests that, under this benchmark, these methods rely more on view-dependent radiance approximations than multi-view consistent geometry. DynOmo's dense spatial regularization and foundation model features maintain coherence across the $\approx$2.6cm and $\approx$18.7° viewpoint shift, achieving near-photorealistic quality, but at substantial computational cost: 16 seconds/frame versus Endo-4DGS's 1 second/frame in training time. This 16$\times$ slowdown exposes a critical open problem: current 3D regularization strategies require expensive neighborhood computation, multi-constraint optimization, and dense feature matching, a computational barrier that remains prohibitive for real-time clinical deployment.

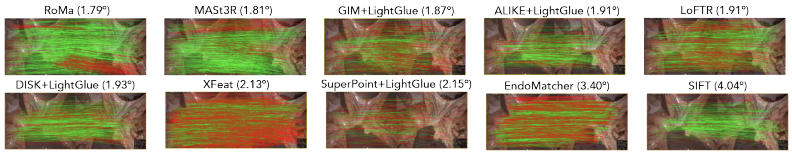

Fig. 7: Qualitative comparison of image matching methods showing correspondence quality. Green lines indicate inlier matches; red lines indicate outliers. The angle next to the method name is the average rotational error.

4.3 Pose Estimation & Feature Matching

Methods. We evaluate 11 pose estimation and feature matching methods on 30 calibrated stereo endoscopy sequences with the reference camera poses. Methods include handcrafted features (SIFT [38]), learned sparse matchers (SuperPoint+LightGlue [13,36], DISK+LightGlue [63], ALIKE+LightGlue [75], XFeat [48]), dense matchers leveraging foundation models (GIM+LightGlue [56], LoFTR [59], EndoMatcher [69], RoMa [15], MASt3R [31]), and a metric relative pose estimator (MicKey [6]). We compute rotation error and inlier rate across all 30 sequences to measure generalization across diverse surgical scenes, anatomies, and imaging conditions. For the 23 sequences featuring tool-tissue deformations with static cameras, where scene motion occurs without camera movement, we additionally compute temporal consistency (standard deviation of rotation errors), which measures frame-to-frame stability critical for tracking and SLAM applications. *Results.* Dense foundation model methods achieve best absolute accuracy: RoMa produces 3905 inliers with 78.1% inlier rate and 1.79° rotation error; MASt3R yields 1676 inliers with 70.3% rate and 1.81° error, providing $2-50\times$ more correspondences than sparse methods. However, learned sparse methods achieve superior temporal consistency: GIM+LightGlue demonstrates 0.41° temporal drift versus $1.38-2.00°$ for dense methods, making it ideal for real-time SLAM despite slightly lower absolute accuracy. This temporal stability likely arises from GIM's label propagation across video frames during training, which enforces match coherence over time [56]. MicKey [6], a metric relative pose estimator trained on outdoor scenes, reasonably failed to produce valid correspondences across all surgical sequences, presenting an opportunity for future work to develop efficient metric pose estimators robust to deformable environments. SIFT's poor performance confirms that learned methods provide better accuracy with lower variance compared to handcrafted features. These results demonstrate that models trained on large-scale diverse data transfer effectively to surgical domains, and that the accuracy-stability tradeoff between dense and sparse methods persists even in constrained imaging conditions.

4.4 Monocular Depth Estimation

Methods. We evaluate monocular depth prediction on calibrated stereo endoscopy sequences using reference depth obtained from a SOTA stereo prior [66]. Given a

Table 5: Pose estimation and feature matching benchmarking results. Temporal consistency measures rot. drift across seqs. with tool-tissue deformation and static cameras.

Method	Rot. Err.(°)↓	Temporal Drift(°)↓	Inlier Rate (%)↑	Avg Inliers↑
RoMa [15]	1.79±1.41	1.41	78.1	3905
MASt3R [31]	1.81±1.38	1.38	70.3	1676
GIM+LG [36, 56]	1.87±1.48	0.41	43.1	264
ALIKE+LG [36, 75]	1.91±1.52	0.64	68.2	211
LoFTR [59]	1.91±1.40	0.43	47.2	513
DISK+LG [36, 63]	1.93±1.52	0.59	55.2	446
XFeat [48]	2.13±1.47	0.77	25.9	854
SP+LG [13, 36]	2.15±1.41	0.70	34.2	74
EndoMatcher [69]	3.40±1.91	2.11	39.9	1255
SIFT [38]	4.04±14.45	5.44	52.9	91

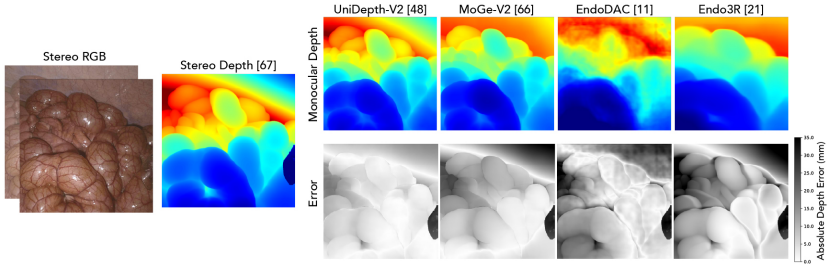

Fig. 8: Qualitative comparison of mono. depth estimation compared to *inferred* stereo depth [66]. Error maps show abs. depth error after median scaling of the predictions.

left image, each model predicts dense depth $\hat{d}_i$, which we compare to the stereo reference d_i reprojected into the same frame. We benchmark four representative models without sequence-specific fine-tuning: two domain-specific surgical models, Endo3R [21] and EndoDAC [11], and two recent metric-scale SOTA monocular depth models, UniDepth-V2 [47] and MoGe-V2 [65]. Evaluation is performed on the same 30 test sequences from ex-vivo, in-vivo, and postmortem scenes (over 6×10^8 valid pixels). We report standard depth metrics: absolute relative error (Abs Rel), root mean squared error (RMSE), log RMSE, and threshold accuracies δ_1, δ_2, δ_3, where $\max(\hat{d}_i/d_i, d_i/\hat{d}_i) < 1.25^k$, $k \in \{1, 2, 3\}$. Since monocular depth is inherently ambiguous in global scale, we report results *with median scaling* (single optimal scale per sequence). For two models that explicitly claim metric-scale prediction [47, 65], we also report *without scaling*.

Results. With median scaling, the two general-purpose, metric-scale monocular depth models [47,65]) outperform the surgical-domain methods [11,21](Tab. 6, Fig. 8). This holds despite EndoDAC being initialized via [70] and subsequently fine-tuned on surgical data, while Endo3R is trained only on surgical imagery. In the no-scaling setting, the two models that claim metric scale exhibit near-zero δ accuracies and large errors. This indicates a substantial global scale mismatch and reinforces that explicit scale calibration is necessary in surgical deployment.

Table 6: Monocular depth prediction compared to *inferred* stereo depth [66]. Methods that report metric-scale predictions [47, 65] are evaluated without median scaling.

Method	Med. Scaled	Abs Rel↓	RMSE↓	$\text{RMSE}_{\log}$↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
UniDepth-V2 [47]	Yes	0.081	0.012	0.111	0.936	0.992	0.998
MoGe-V2 [65]	Yes	0.109	0.014	0.140	0.889	0.989	0.998
EndoDAC [11]	Yes	0.119	0.017	0.164	0.854	0.968	0.994
Endo3R [21]	Yes	0.187	0.023	0.236	0.704	0.927	0.972
UniDepth-V2 [47]	No	11.573	1.098	2.465	0.000	0.000	0.000
MoGe-V2 [65]	No	14.556	1.329	2.683	0.000	0.000	0.000

Limitations. *Ex vivo* sequences use hand-guided (non-robotic) endoscope motion, resulting in less smooth trajectories than robotic manipulation. For dynamic camera sequences where ArUco calibration markers remain visible throughout, mask-based preprocessing is used to exclude markers [34]. Viewpoint diversity is lower than classic multi-view datasets because surgical acquisition is constrained by trocar placement, visibility, and safe camera motion. More subject and anatomical diversity is important future work.

5 Conclusion

iMED exposes limitations in 3D perception under constrained, deformable surgical environments through synchronized recordings from two independent stereo endoscopes with overlapping fields of view. This enables evaluation from physically distinct viewpoints, revealing whether methods recover consistent geometry or mainly interpolate appearance along a single trajectory. Across rigid, deformable, and online novel view synthesis, pose estimation/feature matching, and monocular depth estimation, many methods degrade under narrow-baseline motion, specular tissue, full-field deformation, and topology changes. Surgical scenes therefore act as a stress test for assumptions often hidden by standard same-trajectory evaluations.

Across tasks, broad multi-domain vision models often transfer more robustly to unseen surgical domains than surgical-only models or aggressively fine-tuned methods, suggesting that narrow training can discard invariances needed for cross-device and cross-site generalization. We hypothesize that robust surgical 3D perception will require multi-domain training that incorporates surgical data without sacrificing general visual priors. iMED provides infrastructure for this direction, with diverse anatomical conditions and a benchmark that rewards genuine geometric generalization. Advances under these constraints are likely to transfer to robotics, AR/VR, and embodied systems operating in dynamic, sensor-limited environments. **Acknowledgements.** This work was supported in part by Intuitive Surgical and by the Engineering and Physical Sciences Research Council (EPSRC) through the OASIS programme under grant UKRI145.

References

1. Aanæs, H., Jensen, R.R., Vogiatzis, G., Tola, E., Dahl, A.B.: Large-scale data for multiple-view stereopsis. International Journal of Computer Vision **120**(2), 153–168 (2016)
2. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
3. Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zeffiro, T., Xia, W., et al.: Stereo correspondence and reconstruction of endoscopic data challenge. arXiv preprint arXiv:2101.01133 (2021)
4. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
6. Barroso-Laguna, A., Bokman, G., Edstedt, J., Wadenback, M., Shi, J., Felsberg, M.: Matching 2d images in 3d: Metric relative pose from metric correspondences. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27067–27077 (2024), oral Paper
7. Bertiche, H., Madadi, M., Escalera, S.: Cloth3d: clothed 3d humans. In: European Conference on Computer Vision. pp. 344–359. Springer (2020)
8. Broxton, M., Flynn, J., Overbeck, R., Erickson, D., Hedman, P., Duvall, M., Dourgarian, J., Busch, J., Whalen, M., Debevec, P.: Immersive light field video with a layered mesh representation. ACM Transactions on Graphics (TOG) **39**(4), 86–1 (2020)
9. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
10. Chadebecq, F., Vasconcelos, F., Lacher, R., Maneas, E., Desjardins, A., Ourselin, S., Vercauteren, T., Stoyanov, D.: Refractive two-view reconstruction for underwater 3d vision. International Journal of Computer Vision **128**(5), 1101–1117 (2020)
11. Cui, B., Islam, M., Bai, L., Wang, A., Ren, H.: Endodac: Efficient adapting foundation model for self-supervised depth estimation from any endoscopic camera. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 208–218. Springer (2024)
12. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
13. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 224–236 (2018)
14. Ding, H., Zhang, Y., Lu, T., Liang, R., Shu, H., Seenivasan, L., Long, Y., Dou, Q., Gao, C., Leng, Y., et al.: Segstrong-c: Segmenting surgical tools robustly on non-adversarial generated corruptions–an endovis' 24 challenge. arXiv preprint arXiv:2407.11906 (2024)
15. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19990–19999 (2024)

16. Edwards, P.E., Psychogyios, D., Speidel, S., Maier-Hein, L., Stoyanov, D.: Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. Medical image analysis (2022)
17. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981). https://doi.org/10.1145/358669.358692
18. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Marín-Jiménez, M.J.: Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition **47**(6), 2280–2292 (2014)
19. Garrido-Jurado, S., Munoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R.: Generation of fiducial marker dictionaries using mixed integer linear programming. Pattern recognition **51**, 481–491 (2016)
20. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3354–3361. IEEE (2012)
21. Guo, J., Dong, W., Huang, T., Ding, H., Wang, Z., Kuang, H., Dou, Q., Liu, Y.H.: Endo3r: Unified online reconstruction from dynamic monocular endoscopic video. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 170–180. Springer (2025)
22. Hartwig, R., Ostler, D., Rosenthal, J.C., Feußner, H., Wilhelm, D., Wollherr, D.: Miti: Slam benchmark for laparoscopic surgery. arXiv preprint arXiv:2202.11496 (2022)
23. Hayoz, M., Hahne, C., Kurmann, T., Allan, M., Beldi, G., Candinas, D., Márquez-Neila, P., Sznitman, R.: Online 3d reconstruction and dense tracking in endoscopic videos. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 444–454. Springer (2024)
24. Huang, Y., Cui, B., Bai, L., Guo, Z., Xu, M., Islam, M., Ren, H.: Endo-4dgs: Endoscopic monocular scene reconstruction with 4d gaussian splatting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 197–207. Springer (2024)
25. Hwang, H., Lim, J., Kinnaird, C., Nagy, A., Panton, O., Hodgson, A., Qayumi, K.: Correlating motor performance with surgical error in laparoscopic cholecystectomy. Surgical Endoscopy And Other Interventional Techniques **20**(4), 651–655 (2006)
26. Itseez: Open source computer vision library. https://github.com/itseez/opencv (2015)
27. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: The IEEE International Conference on Computer Vision (ICCV) (2015)
28. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
30. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)
31. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European Conference on Computer Vision (ECCV). pp. 74–92. Springer (2024)

32. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quarterly of applied mathematics **2**(2), 164–168 (1944)
33. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5521–5531 (2022)
34. Li, X., Xue, H., Ren, P., Bo, L.: Diffueraser: A diffusion model for video inpainting. arXiv preprint arXiv:2501.10018 (2025)
35. Limperg, T., Novoa, V., Curlin, H., Veersema, S.: Laparoscopic trocar dimensions: Marketed versus true dimensions–a descriptive study. Journal of Minimally Invasive Gynecology **28**(11), S79 (2021)
36. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17627–17638 (2023)
37. Liu, X., Tayal, P., Wang, J., Zarzar, J., Monnier, T., Tertikas, K., Duan, J., Toisoul, A., Zhang, J.Y., Neverova, N., et al.: Uncommon objects in 3d. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14102–14113 (2025)
38. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
39. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameter. Journal of the society for Industrial and Applied Mathematics **11**, 431–441 (1963)
40. Michel Hayoz, Christopher Hahne, M.G.D.C.T.K.M.A.R.S.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery 2023 . `https://doi.org/https://doi.org/10.1007/s11548-023-02919-w`
41. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**, 99–106 (2021)
42. Mountney, P., Stoyanov, D., Yang, G.Z.: Three-dimensional tissue deformation recovery and tracking. IEEE Signal Processing Magazine **27**(4), 14–24 (2010)
43. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d camera. IEEE transactions on robotics **33**, 1255–1262 (2017)
44. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021)
45. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021)
46. Penza, V., Ciullo, A.S., Moccia, S., Mattos, L.S., De Momi, E.: Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms. The International Journal of Medical Robotics and Computer Assisted Surgery **14**(5), e1926 (2018)
47. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
48. Potje, G., Cadar, F., Araujo, A., Martins, R., Nascimento, E.R.: Xfeat: Accelerated features for lightweight image matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2682–2691 (2024)

49. Pratt, P., Stoyanov, D., Visentini-Scarzanella, M., Yang, G.Z.: Dynamic guidance for robotic surgery using image-constrained biomechanical models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–85. Springer (2010)
50. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10318–10327 (2021)
51. Recasens Lafuente, D., Oswald, M.R., Pollefeys, M., Civera, J.: The drunkard's odometry: Estimating camera motion in deforming scenes. Advances in Neural Information Processing Systems **36**, 48877–48889 (2023)
52. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: 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. pp. 10901–10911 (2021)
53. Schmidt, A., Mohareri, O., DiMaio, S.P., Salcudean, S.E.: Surgical tattoos in infrared: A dataset for quantifying tissue tracking and mapping. IEEE Transactions on Medical Imaging **43**(7), 2634–2645 (2024)
54. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
55. Seidenschwarz, J., Zhou, Q., Duisterhof, B.P., Ramanan, D., Leal-Taixé, L.: Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. In: 2025 International Conference on 3D Vision (3DV). pp. 1012–1021. IEEE (2025)
56. Shen, X., Jiang, Z., Huang, R., Stojanov, S., Kumar, A., Xiang, K., Wang, S., Bao, H., Wang, Y., Wu, C., et al.: Gim: Learning generalizable image matcher from internet videos. In: The Twelfth International Conference on Learning Representations (ICLR) (2024), spotlight Paper
57. Shoemake, K.: Animating rotation with quaternion curves. In: Proceedings of the 12th annual conference on Computer graphics and interactive techniques. pp. 245–254 (1985)
58. Stoyanov, D., Scarzanella, M.V., Pratt, P., Yang, G.Z.: Real-time stereo reconstruction in robotically assisted minimally invasive surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 275–282. Springer (2010)
59. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8922–8931 (2021)
60. Supe, A.N., Kulkarni, G.V., Supe, P.A.: Ergonomics in laparoscopic surgery. Journal of minimal access surgery **6**(2), 31–36 (2010)
61. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: European conference on computer vision. pp. 402–419. Springer (2020)
62. Turki, H., Agrawal, V., Bulò, S.R., Porzi, L., Kontschieder, P., Ramanan, D., Zollhöfer, M., Richardt, C.: Hybridnerf: Efficient neural rendering via adaptive volumetric surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19647–19656 (2024)
63. Tyszkiewicz, M., Fua, P., Trulls, E.: Disk: Learning local features with policy gradient. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 14254–14265 (2020)
64. Wang, Q., Khanicheh, A., Leiner, D., Shafer, D., Zobel, J.: Endoscope field of view measurement. Biomedical optics express **8**(3), 1441–1454 (2017)

65. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025)
66. Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S.: Foundationstereo: Zero-shot stereo matching. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5249–5260 (2025)
67. Wickrema, C., Leary, S., Sarkar, S., Giglio, M., Bianchi, E., Mace, E., Twardowski, M.: Benchmarking image similarity metrics for novel view synthesis applications. arXiv preprint arXiv:2506.12563 (2025)
68. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
69. Yang, B., Tian, Q., Geng, Y., Liao, H., Huang, X., Luo, J., Liu, H.: Endomatcher: Generalizable endoscopic image matcher via multi-domain pre-training for robot-assisted surgery. arXiv preprint arXiv:2508.05205 (2025)
70. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
71. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)
72. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
73. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1790–1799 (2020)
74. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
75. Zhao, X., Wu, X., Chen, W., Chen, P.C., Xu, Q., Li, Z.: Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. IEEE Transactions on Instrumentation and Measurement **72**, 1–16 (2023)
76. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)