2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

ROOM: A Physics-Based Continuum Robot Simulator for Photorealistic Medical Datasets Generation

Salvatore Esposito[1], Matías Mattamala[1], Daniel Rebain[2], Francis Xiatian Zhang[1], Kevin Dhaliwal[1], Mohsen Khadem[1], and Subramanian Ramamoorthy[1]

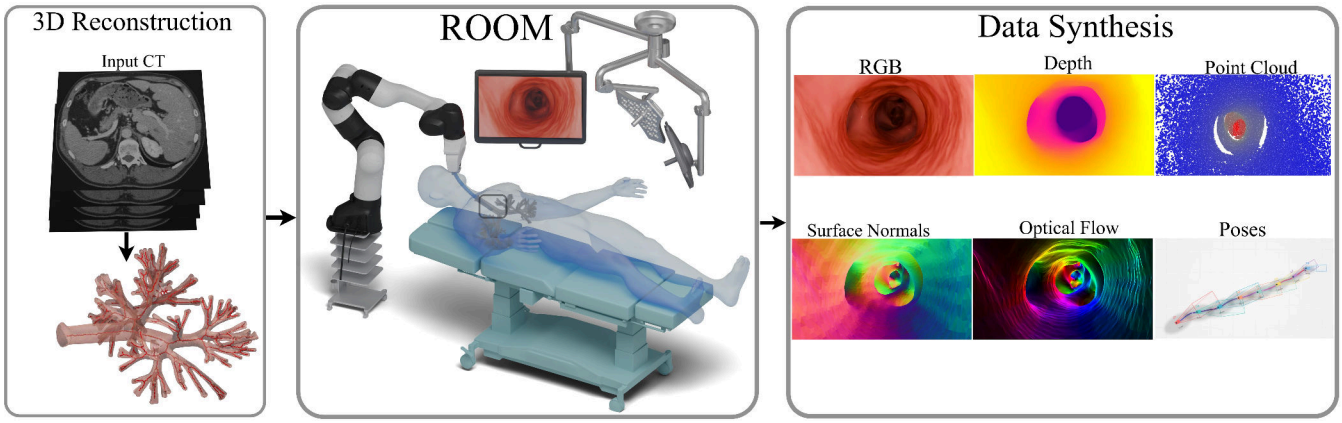

Fig. 1: **ROOM framework overview**. Given patient CT scans (left), our pipeline reconstructs accurate 3D lung models and extracts medial axis trajectories, enabling physics-based continuum robot simulation to generate photorealistic multi-modal sensor data (right). This includes RGB images with realistic noise and lighting, metric depth maps, surface normals, optical flow, point clouds, and ground-truth poses, for different medical robotics applications.

***Abstract*— Continuum robots are advancing bronchoscopy procedures by accessing complex lung airways and enabling targeted interventions. However, their development is limited by the lack of realistic training and test environments: Real data is difficult to collect due to ethical constraints and patient safety concerns, and developing autonomy algorithms requires realistic imaging and physical feedback. We present ROOM (Realistic Optical Observation in Medicine), a comprehensive simulation framework designed for generating photorealistic bronchoscopy training data. By leveraging patient CT scans, our pipeline renders multi-modal sensor data including RGB images with realistic noise and light specularities, metric depth maps, surface normals, optical flow and point clouds at medically relevant scales. We validate the data generated by ROOM in two canonical tasks for medical robotics: multi-view pose estimation and monocular depth estimation, demonstrating diverse challenges that state-of-the-art methods must overcome to transfer to these medical settings. Furthermore, we show that the data produced by ROOM can be used to fine-tune existing depth estimation models to overcome these challenges, also enabling other downstream applications such as navigation. We expect that ROOM will enable large-scale data generation across diverse patient anatomies and procedural scenarios that are challenging to capture in clinical settings. Code and data: https://iamsalvatore.io/room/.**

I. INTRODUCTION

Continuum robots have emerged as an innovative technology in minimally invasive surgery, with bronchoscopy representing one of the most promising applications. These flexible, cable-driven systems can navigate the intricate branching networks of human airways with unprecedented dexterity, enabling precise drug delivery, tissue sampling, and diagnostic imaging in lung regions previously inaccessible to rigid instruments [1]. Continuum robots can enable early intervention in peripheral lung nodules, targeted chemotherapy delivery, and real-time biopsy guidance, significantly improving patient outcomes in pulmonary medicine.

Nevertheless, the development of autonomous navigation algorithms for continuum robot bronchoscopy faces data-related limitations. Clinical data collection is inherently constrained by patient safety protocols, ethical review processes, and the high costs associated with experimental procedures. More fundamentally, the individualised nature of human anatomy means that effective algorithms must generalise across diverse airway geometries while maintaining millimetre-level precision [1]. Synthetic data generation has demonstrated remarkable success in addressing similar challenges across robotics applications from autonomous driving to visual SLAM [2], [3]. In the medical context, some recent efforts have focused on data generation for colonoscopy, as done by the SimCol3D Challenge [4], or on simulation frameworks for surgical procedures [5]. However, bronchoscopy procedures require anatomical fidelity, procedure-specific lighting conditions, as well as specific kinematics and sensor modalities calibrated to clinical scales.

In this paper, we introduce ROOM (Realistic Optical Observation in Medicine), a simulation framework engineered for continuum robot bronchoscopy applications. ROOM

[1]University of Edinburgh, UK. [2]University of British Columbia, Canada.
This work was supported by a UKRI Turing AI World Leading Researcher Fellowship on AI for Person-Centred and Teachable Autonomy (grant EP/Z534833/1)

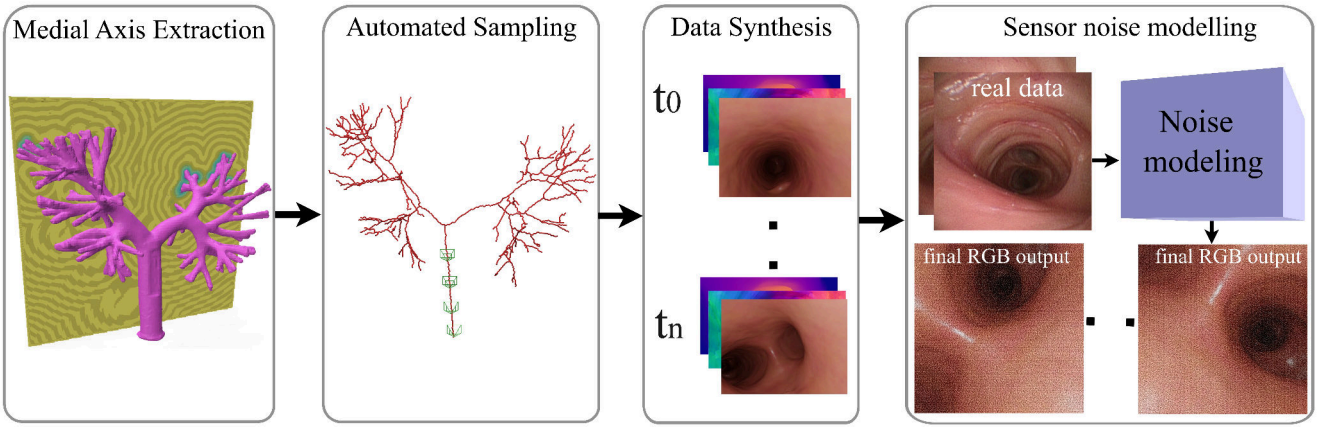

Fig. 2: **ROOM data generation pipeline**. The system consists of four main stages: (1) Medial Axis Extraction from segmented CT lung models, (2) Automated Sampling along skeletal branches with higher density at bifurcations and high-curvature regions, (3) Data Synthesis generating synchronized multi-modal sensor streams from t_0 to t_n timesteps, and (4) Sensor Noise Modeling applying realistic noise characteristics matching real bronchoscopy imagery through frequency-domain analysis.

provides the first fully automated pipeline that transforms patient CT scan data into extensive synthetic training datasets while preserving the geometric constraints and visual characteristics essential for medical navigation tasks inside the vessels and airways of the anatomical structures. Our system generates photorealistic multi-modal sensor data, including RGB imagery with realistic noise, metric depth maps, surface normals, point clouds, and optical flow, all calibrated to the millimetre scales typical of bronchoscopy procedures, as shown in Fig. 2. By enabling large-scale data generation across diverse patient anatomies and challenging procedural scenarios, ROOM can facilitate the development of robot bronchoscopy without the constraints of clinical data collection. The primary contributions of this work are:

- ROOM, a realistic simulation framework designed for continuum robot bronchoscopy to generate synthetic data at medically-relevant scales.
- A photorealistic rendering pipeline that considers endoscopic lighting conditions, tissue surface properties, and data-driven sensor models.
- Validation of the synthetic data produced by ROOM in medically-relevant tasks, such as multi-view pose estimation and monocular depth estimation.
- Demonstration of additional applications such as monocular depth fine-tuning and visual navigation.
- Open-source release of ROOM for benefit of the community at https://iamsalvatore.io/room/.

II. Related Work

Medical Robotics Simulators. Specialised simulation platforms for medical robotics have primarily targeted surgical training and haptic feedback [6], [5], [7], offering real-time interaction but simplified visual rendering insufficient for training sim-to-real vision systems [8], [9]. Recent neural rendering and GPU-accelerated platforms such as ORBIT-Surgical [5] achieve fast real-time rendering for surgery and endoscopy simulation [10], [11], yet are not designed for large-scale dataset generation or multi-modal sensor output (depth, optical flow, surface normals) required for navigation and depth estimation. In colonoscopy, SimCol3D [4] introduced a Unity-based synthetic data pipeline for 3D reconstruction, pose estimation, and monocular depth estimation. However, bronchoscopy presents additional appearance and geometric degeneracies compared to the texture- and geometry-rich colon environment, necessitating advanced rendering techniques such as path tracing and BSDF shaders that ROOM integrates within its pipeline.

Continuum Robot Bronchoscopy Systems. Continuum robots have shown significant potential in bronchoscopy, with clinical studies demonstrating improved diagnostic accuracy through flexible navigation of complex airway geometries [1], [12]. Prior efforts have focused on odometry and localisation: PANS [13] demonstrated 6-DOF pose tracking without external sensors via Monte-Carlo localisation given a prior lung map, while Deng et al. [14] introduced an ex-vivo dataset for evaluating monocular visual odometry in map-free settings. While we do not target these specific tasks, we show how ROOM-generated data supports multi-view pose estimation. The ultimate goal of continuum bronchoscopy robots is autonomous (or semi-autonomous) navigation for localised procedures. Prior work has acquired reference trajectories in simulation [15] or from real data [16], [17], while more recent approaches learn navigation policies via reinforcement learning in simulation [18], [19]. However, these decouple physics simulation from photorealism, limiting policy performance. ROOM bridges this gap with a unified framework for visually-accurate data collection in physically-realistic settings.

III. Method

A. Overview

ROOM provides a comprehensive simulation framework for generating photorealistic bronchoscopy training data using continuum robots. The system comprises four components: (1) continuum robot modelling with realistic kinematic constraints, (2) physics simulation with calibrated tissue interactions, (3) anatomical reconstruction and trajectory

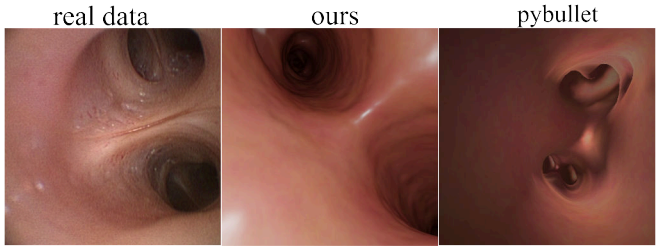

Fig. 3: **Visual comparison of ROOM outputs compared to real data.** *Left:* Real bronchoscopy data captured from a continuum robot showing specular highlights from wet mucosal surfaces and directional lighting. *Center:* ROOM's photorealistic rendering using Blender's path tracing with Principled BSDF shaders, accurately reproducing tissue surface properties and lighting conditions. *Right:* Naive PyBullet-based rendering lacking photorealistic materials and lighting.

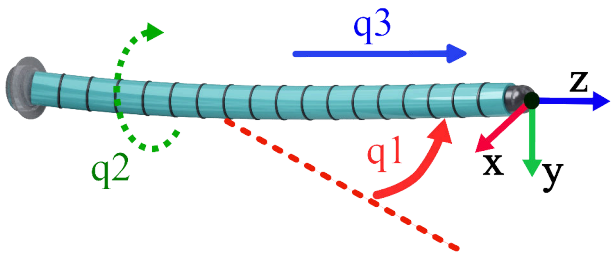

Fig. 4: **Continuum robot model used in ROOM simulation**. The bronchoscope is modelled as a flexible, cable-driven continuum robot with constant curvature bending and three degrees of freedom: tendon actuation for bending curvature (q1), axial rotation for bending plane (q2), and linear insertion depth (q3). The physics-based simulation incorporates realistic friction models, actuator noise, and collision dynamics calibrated to clinical bronchoscope behaviour.

planning, and (4) photorealistic rendering with endoscopic artifacts. We describe these components below.

B. Continuum Robot Modelling

The bronchoscope is modelled as a cable-driven continuum robot using a reduced-order piecewise-constant-strain approximation (Fig. 4). The model exposes three DoF aligned with clinical control interfaces: antagonistic tendon differential displacement for bending ($q_1 \in [-0.008, 0.008]$ m), axial rotation selecting the bending plane ($q_2 \in \mathbb{R}$ rad), and linear insertion depth ($q_3 \in \mathbb{R}$ m).

Cosserat rod kinematics. The backbone configuration along arc-length $s \in [0, l]$ is described by centreline position $r(s) \in \mathbb{R}^3$ and material frame $R(s) \in SO(3)$. In Cosserat form,

$$\frac{dr(s)}{ds} = R(s)\,v(s), \qquad \frac{dR(s)}{ds} = R(s)\,\widehat{u}(s), \tag{1}$$

where $v(s) \in \mathbb{R}^3$ is the translational strain, $u(s) = [u_x, u_y, u_z]^\top$ is the body-frame curvature–torsion strain, and $\widehat{(\cdot)}$ denotes the skew-symmetric operator. Assuming an inextensible, unshearable section gives the Kirchhoff reduction $v(s) = e_3 = [0, 0, 1]^\top$, hence $dr/ds = R(s)e_3$. In the full Cosserat formulation, $u(s)$ follows from equilibrium and a constitutive law; here it is used only to define the nominal unloaded shape.

Boundary conditions and constant-curvature actuation. The base boundary conditions at $s = 0$ are

$$r(0) = \begin{bmatrix} 0 & 0 & q_3 \end{bmatrix}^\top, \qquad R(0) = \mathrm{Rot}_z(q_2), \tag{2}$$

where q_3 sets insertion depth and q_2 rotates the bending plane. Assuming constant strain $u(s) = u_0$ along the distal flexible segment:

$$u_0 = \begin{bmatrix} \kappa(q_1) & 0 & 0 \end{bmatrix}^\top, \qquad \kappa(q_1) = -\frac{q_1}{2\gamma\, l}\ \ [\mathrm{m}^{-1}], \tag{3}$$

where $l = 50 \times 10^{-3}$ m is the flexible segment length and $\gamma = 1.75 \times 10^{-3}$ m is the tendon routing radius. For a symmetric antagonistic tendon pair, this follows from the small-curvature relation $q_1 \approx -2\gamma l\kappa$; axial strain and torsion are neglected.

Physics-based simulation and interaction modelling. The robot is simulated in PyBullet via a compliant discretisation of the backbone into capsule collision bodies connected by revolute joints. Joint targets encode the nominal constant-curvature shape in Eq. 3, while joint stiffness $k_j = 2.0 \times 10^{-2}$ N m rad^{-1} and damping $d_j = 5.0 \times 10^{-4}$ N m s rad^{-1} provide compliance under external loads. External contact and friction are not imposed as boundary conditions of the reduced-order rod model; instead, they are resolved by the compliant multibody simulation about the nominal shape.

Friction and soft contact. Bronchoscope–airway contact follows a Coulomb friction model with $\mu_s = 0.3$ and $\mu_d = 0.25$. Normal interaction is modelled by

$$f_n = \max\Big(0,\ k_n\,\delta + c_n\,\dot{\delta}\Big), \tag{4}$$

where $\delta \geq 0$ is penetration depth, $\dot{\delta}$ the relative normal velocity, and $k_n = 4.0 \times 10^3$ N m^{-1}, $c_n = 0.8$ N s m^{-1}. These values correspond to $\delta = 0.25$ mm at 1 N and $\delta = 0.5$ mm at 2 N, thereby keeping nominal wall penetration sub-millimetre over the relevant load range.

Actuation non-idealities. To capture transmission delay and control non-idealities, we inject stochastic control delays $\Delta t \sim \mathcal{U}(0, 0.1)$ s and magnitude-dependent scaling terms:

$$q_1(t) = q_{1,\mathrm{cmd}}(t - \Delta t)\left(1 + 0.05\,\frac{|q_{1,\mathrm{cmd}}(t - \Delta t)|}{q_{1,\max}}\right), \tag{5}$$

$$q_2(t) = q_{2,\mathrm{cmd}}(t - \Delta t)\left(1 + 0.05\,\frac{|q_{2,\mathrm{cmd}}(t - \Delta t)|}{2\pi}\right). \tag{6}$$

C. Anatomical Reconstruction and Data Synthesis

CT Scan Preprocessing. Patient-specific anatomical models are extracted from clinical CT scans through an automated segmentation-to-mesh pipeline. We first resample each CT volume to an isotropic grid and apply standard intensity normalisation for lung CT (HU windowing followed by affine normalisation). Airway lumen segmentation is produced using a 3D U-Net-style based nnUNet encoder-decoder with skip connections [20]. Training is patch-based with sampling biased toward airway regions and uses geometric and intensity augmentations (random rotations/scales and intensity jitter).

Automated Data Collection. To automatically collect data within the airways, we generate collision-free trajectories

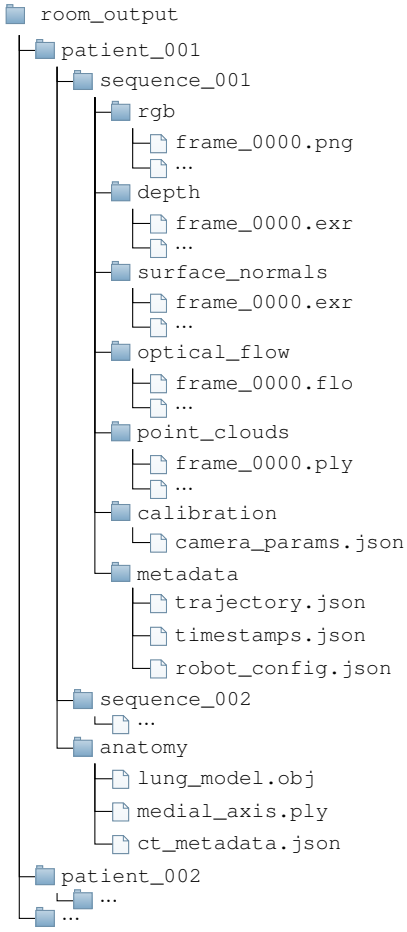

Fig. 5: **ROOM pipeline output folder structure.** The framework generates synchronized multi-modal sensor data organized by patient anatomy and sequence. Each sequence contains RGB images (600×600), metric depth maps, surface normals, optical flow fields, point clouds, ground-truth poses, and calibration parameters with timestamps.

by extracting the medial axis from the reconstructed signed distance field (SDF) of the airway geometry (Fig. 2). Starting from surface sample points, we trace along the SDF gradient $\nabla\phi(\mathbf{x})$ in the inward normal direction. Following the grass-fire analogy [21], trajectories $\mathbf{x}(t)$ propagate inward with $\dot{\mathbf{x}}(t) = -\mathbf{n}(t)$. Medial axis points are identified where the gradient exhibits sign changes, i.e. $\frac{d}{dt}[\nabla\phi(\mathbf{x}(t))] \cdot \hat{\mathbf{n}} = 0$, detected via pronounced spikes in the second derivative $\nabla^2\phi(\mathbf{x})$. The extracted medial axis forms a navigation graph capturing the airway centreline topology. Trajectory sampling along this skeleton uses adaptive density at bifurcations and high-curvature regions to ensure comprehensive coverage of geometrically complex areas. These medial axis poses serve as collision-free waypoints tracked by an inverse kinematics controller, producing target 6-DoF poses sampled at 10 Hz for rendering photorealistic data streams.

Multi-Modal Data Rendering. For each target pose, we synthesise synchronised data streams: RGB images (600×600), metric and relative depth maps, surface normals, optical flow fields, and point clouds. Each frame is stored together with camera intrinsics/extrinsics, timestamps, and robot configuration to ensure temporal alignment across modalities. Fig. 5 summarises the reconstruction, simulation, and rendered outputs (modalities and metadata). The rendering pipeline utilises Blender's Principled BSDF shader system with physically-based material properties (base colour, metallic, roughness) to reproduce tissue appearance. We model the directional lighting of the bronchoscope by attaching a point light source with exponential falloff to the tip. For non-RGB modalities we use Blender multi-pass rendering: depth is extracted via the Z-buffer, surface normals are computed from geometry derivatives, and optical flow is computed from inter-frame motion vectors.

Method	RRA@5° ↑	RTA@5° ↑	AUC@30° ↑
COLMAP [22]	41.00	0.07	16.91
ORB-SLAM3 [23]	71.67	0.17	42.74
DUSt3R [24]	63.00	0.21	54.90
VGGT [25]	79.00	0.25	69.09

TABLE I: **Comparison of methods across five sequences (Seq0–Seq4)**. Reported values are means across all sequences. Metrics: Relative Rotation Accuracy (RRA@5°), Relative Translation Accuracy (RTA@5°), and Area Under the Curve (AUC@30°). Higher is better (↑).

Sensor Noise Modelling. Finally, to accurately reproduce noise characteristics of real bronchoscopy RGB images, we employ a frequency-domain system identification approach. Given real endoscopic data I_{real}, we extract the noise component through bilateral filtering as shown in Fig. 2:

$$n_{\text{real}} = I_{\text{real}} - \text{BF}(I_{\text{real}}) \tag{7}$$

We then analyse the noise spectrum through its Fourier transform $N_{\text{real}}(\omega) = \mathcal{F}\{n_{\text{real}}\}$ and characterize the frequency distribution by the power spectral density $P(\omega) = |N_{\text{real}}(\omega)|^2$. For synthetic data generation, we shape the white noise w to match this spectrum:

$$n_{\text{synth}} = \mathcal{F}^{-1}\left\{\mathcal{F}\{w\} \cdot \sqrt{P(\omega)}\right\} \tag{8}$$

The final synthetic RGB image combines the rendered output with the synthesised noise: $I_{\text{synth}} = I_{\text{rendered}} + \beta \cdot n_{\text{synth}}$, where β controls the noise amplitude to match medical sensor characteristics. This approach ensures our synthetic data exhibits the same noise statistics as real bronchoscopy imagery, which we observed is crucial for assessing monocular depth estimation performance.

IV. Applications

We demonstrate ROOM's data for two canonical tasks in medical robotics: multi-view pose estimation and monocular depth estimation evaluation. Additionally, we demonstrate applications of the synthesised data for fine-tuning depth estimation models, as well as potential navigation tasks.

A. Task 1: Multi-View Pose Estimation

The first task is camera pose estimation from multiple views, a fundamental task in medical robotics that underpins downstream bronchoscopy use-cases such as 3D reconstruction. The repetitive branching patterns and limited texture

of airways, pose particular challenges for evaluating existing visual odometry and structure-from-motion methods.

For evaluation, we synthesised realistic reference paths along the airways, to obtain photorealistic RGB images and ground truth poses. We evaluated four methods: ORB-SLAM [23] as a classical feature-based baseline, COLMAP [22] with sequential matching constraints, and DUSt3R [24] and VGGT [25] as learning-based methods. We measure the Relative Rotation Accuracy (RRA@5°), Relative Translation Accuracy (RTA@5°), and Area Under the Curve (AUC@30°), as done in prior work [25].

Our results are reported in Tab. I. We report that classical methods achieve only 41% RRA and 0.07% RTA tracking success due to insufficient texture, while DUSt3R trained on natural image data reaches 63% RRA and 0.37% RTA on held-out sequences. VGGT demonstrates superior performance with 79% RRA and 0.5% RTA, representing a substantial improvement over classical approaches. These results align with recent findings in endoscopic domains: ORB-SLAM3 achieves only 25% frame localisation success on real colonoscopy sequences [26], while other methods such as CudaSIFT-SLAM shows 70% improvement over ORB-SLAM3 in colonoscopy mapping coverage [27]. Similarly, pose estimation studies in endoscopy report challenges with classical methods, with specialised endoscopic pose estimation achieving errors of 1.43 mm in bronchoscopy and 3.64 mm in colonoscopy [28]. The higher performance of VGGT on our bronchoscopy data is consistent with its demonstrated advantages over DUSt3R and traditional methods across multiple benchmarks [25].

B. Task 2: Monocular Depth Estimation

Monocular depth estimation is an important task in medical robotics [4], since it is challenging to use stereo setups or structured light under the bronchoscope's size constraints—they range from 2.4–6.2 mm in outer diameter [29].

We compare different pre-trained depth estimation models using ROOM-generated data. We evaluate four general-purpose foundation models for monocular depth, namely Metric3D-V2 [30], Depth Anything V2 (monocular and relative variants) [31], and UniDepth (monocular and relative variants) [32]. Additionally, we evaluate three endoscopy-specialized methods: EndoDAC (transfer from Depth Anything) [33], EndoOmni (transfer from DINOv2) [34], and BREA-Depth (transfer from Depth Anything V2) [35]. Each model is evaluated using standard depth estimation metrics: L1 error, RMSE, absolute relative error, and delta accuracy thresholds ($\delta < 1.25^i$ for $i \in \{1, 2, 3\}$).

The quantitative results in Tab. II highlight a persistent domain gap for monocular depth in bronchoscopy. Absolute errors remain on the order of centimetres (e.g., L1 $\approx$ 9.0–11.0 mm), while relative performance is weak across all baselines (Abs Rel $\approx$ 0.42–0.49 and $\delta_1 \approx$ 27–31%), far below the 80–90% δ_1 commonly reported on natural-image benchmarks. Among all methods, UniDepth achieves the lowest absolute error (L1 = 10.6 mm, RMSE = 16.6 mm),

TABLE II: **Monocular depth estimation results on ROOM synthetic data**. L1 and RMSE are reported in millimetres (mm). δ_1 is the percentage of pixels satisfying $\max(\hat{d}/d, d/\hat{d}) < 1.25$. † Relative-depth methods are scale-aligned per sequence using the ground-truth median depth.

Method	L1 ↓	Abs Rel↓	RMSE ↓	δ_1 (%)↑
Metric3DV2 [30]	9.5	0.440	14.5	27.5
DAV2 (Metric) [31]	9.7	0.459	14.7	28.5
DAV2 (Rel.)† [31]	11.3	0.486	17.9	28.2
EndoDAC [33]	9.4	0.432	14.4	29.6
UniDepth [32]	10.6	0.476	16.6	27.1
EndoOmni [34]	9.2	0.428	14.2	30.2
BREA-Depth [35]	9.1	0.421	14.1	30.8

TABLE III: **Comparison of original and fine-tuned models on an external bronchoscope dataset. Bold** indicates improvements over the original model after fine-tuning. L1 and RMSE are in millimetres.

	Method	L1 ↓	Abs Rel ↓	RMSE ↓	δ_1 (%) ↑
Original	UniDepth [32]	8.0	0.545	10.0	19.77
	DAV2 [31]	20.0	0.382	24.0	42.15
	BREA-D [35]	14.0	0.197	19.0	65.39
Fine-tuned	UniDepth [32]	**4.0**	**0.277**	**6.0**	**59.87**
	DAV2 [31]	**15.0**	**0.291**	**20.0**	**55.42**
	BREA-D [35]	**13.0**	**0.192**	**18.0**	**67.70**

while BREA-Depth achieves the best overall relative accuracy (Abs Rel = 0.421, δ_1 = 30.8%), consistent with improved lumen localisation. These results reflect the combined challenges of limited texture, specularities, and extreme depth ranges (2–50 mm) in the bronchoscopy domain.

The error maps in Fig. 6 expose systematic failure modes. Errors cluster at specular highlights where wet mucosal surfaces create photometric inconsistencies, and at geometric discontinuities including bifurcations where the tubular structure transitions. Furthermore, DAV2 variants show more diffuse error patterns, while Metric3DV2 and UniDepth maintain better structural coherence but fail at boundaries. The repetitive branching geometry provides insufficient texture gradients for reliable depth cues, particularly evident in the uniform error distribution across smooth airway walls.

We conclude that the bronchoscopy environment violates core assumptions of existing depth estimation methods, requiring domain-specific training approaches like ROOM's synthetic data generation to bridge this performance gap.

C. Task 3: Fine-tuning Monocular Depth Models

After the findings of Task 2, we proposed to assess if fine-tuning monocular depth models using synthetic ROOM data could provide performance boosts. For this task we used three models: the general-use UniDepth and DepthAnything V2 (DAV2), as well as the bronchoscopy-specialised BREA-D. Furthermore, to avoid testing the models using a test set within the same data distribution of the fine-tuning data, we evaluated them on an external bronchoscope dataset with phantom-based depth ground truth [36]. We compare their performance before and after fine-tuning using the same depth estimation metrics used in Task 2.

We report the results for the different metrics in Tab. III using ten selected representative image-depth pairs. Our results indicate that fine-tuning on synthetic ROOM-generated data improves BREA-Depth across all metrics, with δ_1 ac-

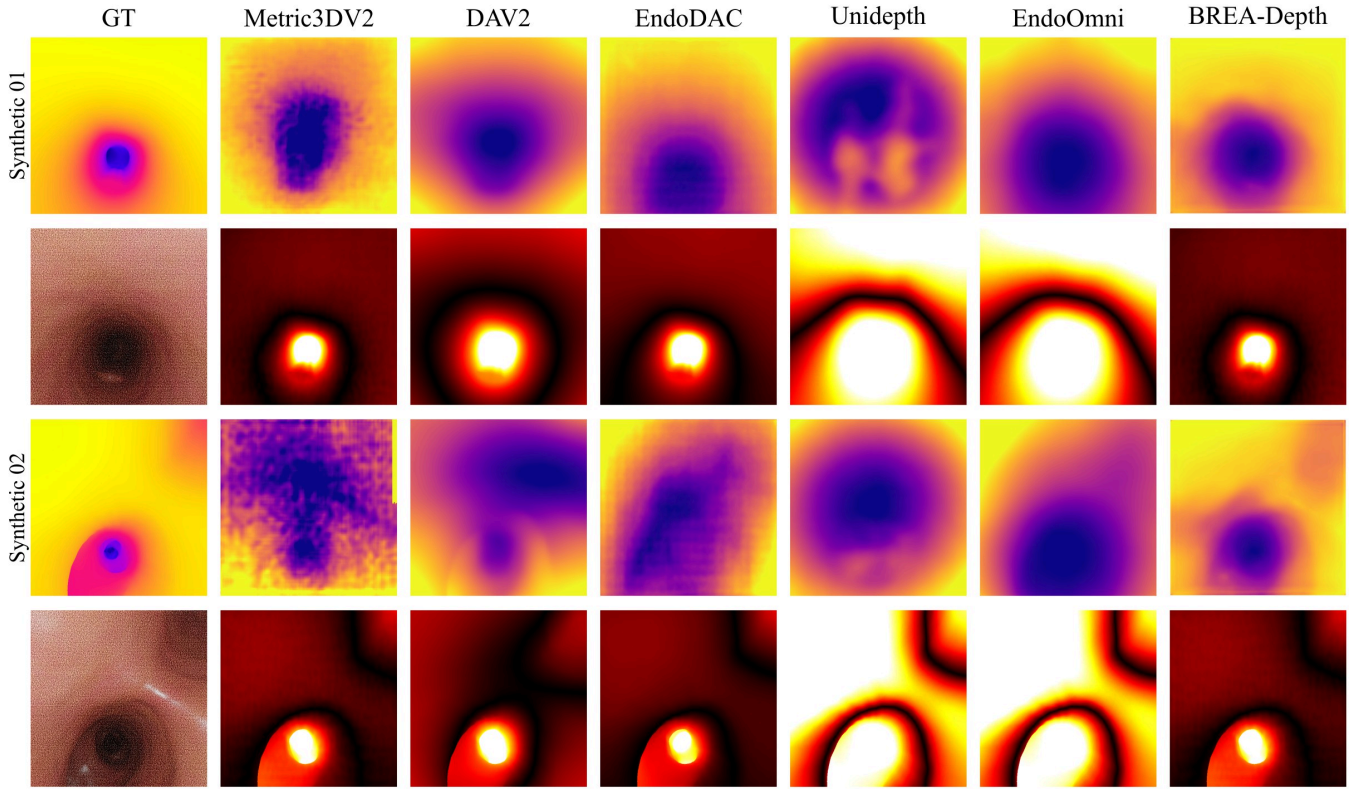

Fig. 6: **Comparative monocular depth estimation results on ROOM synthetic bronchoscopy sequences.** Top rows show L1 error maps between predicted depth estimation and ground truth depth, where warmer colours indicate higher absolute errors, while bottom rows display corresponding RGB inputs with challenging specular highlights and limited texture. Five state-of-the-art models are evaluated: Metric3DV2, Depth Anything V2 (DAV2 Monocular/Relative), Unidepth, EndoOmni, EndoDAC, BREA-Depth, revealing significant performance degradations due to the realistic sensor noise from the simulator and systematic errors concentrated at geometric transitions and specular regions.

curacy increasing from 65.39% to 67.70% (a relative gain of 3.5%). These improvements are also reflected in qualitative comparisons between the pre-trained and fine-tuned models shown in Fig. 7. We report improvements in the fine-tuned models even when tested on real bronchoscopy images that were part of neither the pre-training nor fine-tuning data.

Our results demonstrate that the synthetic data produced by ROOM provides effective supervision for bridging domain gaps and recovering performance under challenging bronchoscopic conditions, suggesting promising avenues to fine-tune general models in this medical domain.

We evaluate ROOM's data on two canonical tasks in medical robotics: multi-view pose estimation and monocular depth estimation. We then use the synthetic data for fine-tuning depth estimation models and report transfer to an external bronchoscopy dataset. Finally, we include a qualitative navigation demonstration to illustrate how ROOM's synchronised modalities (RGB/depth and calibration/poses) can be integrated into a classical planning stack; this demonstration is not presented as a quantitative navigation benchmark.

D. Demonstration: Vision-Based Navigation

Lastly, we provide a qualitative demonstration of how ROOM outputs can be integrated into a vision-based bronchoscope navigation pipeline. The goal is to show that the simulator provides the required synchronised modalities (RGB, depth, and calibration/poses) for downstream planning and control. This is not a quantitative navigation evaluation; therefore do not report navigation success metrics.

We implemented a vision-based navigation method based on a sampling-based planner [37], using the predicted depth maps to generate a local point cloud map for collision checking. Fig. 8 shows example output paths predicted from single frames, visualising the path from the current camera pose (image centre) to the farthest visible point. The 3D visualisations on the right place the planned path in relation to the robot's collision geometry (coloured spheres), illustrating feasibility within lumen constraints. This integration shows that models trained or fine-tuned with ROOM data can be used within classical planning stacks.

V. Discussion

Our results show that the synthetic data produced by ROOM can contribute to overcome challenges that well-established methods in multi-view pose estimation and monocular depth estimation face in the bronchoscopy domain. However, there are limitations and aspects for future improvement of the ROOM framework.

First, the anatomical reconstruction pipeline depends on CT scan quality, and may fail with pathological cases exhibiting severe occlusions or abnormal geometries. However, this also presents an opportunity to extend the framework to other endoscopic procedures where CT scans are available, such as colonoscopy and arthroscopy. Second, while ROOM is built

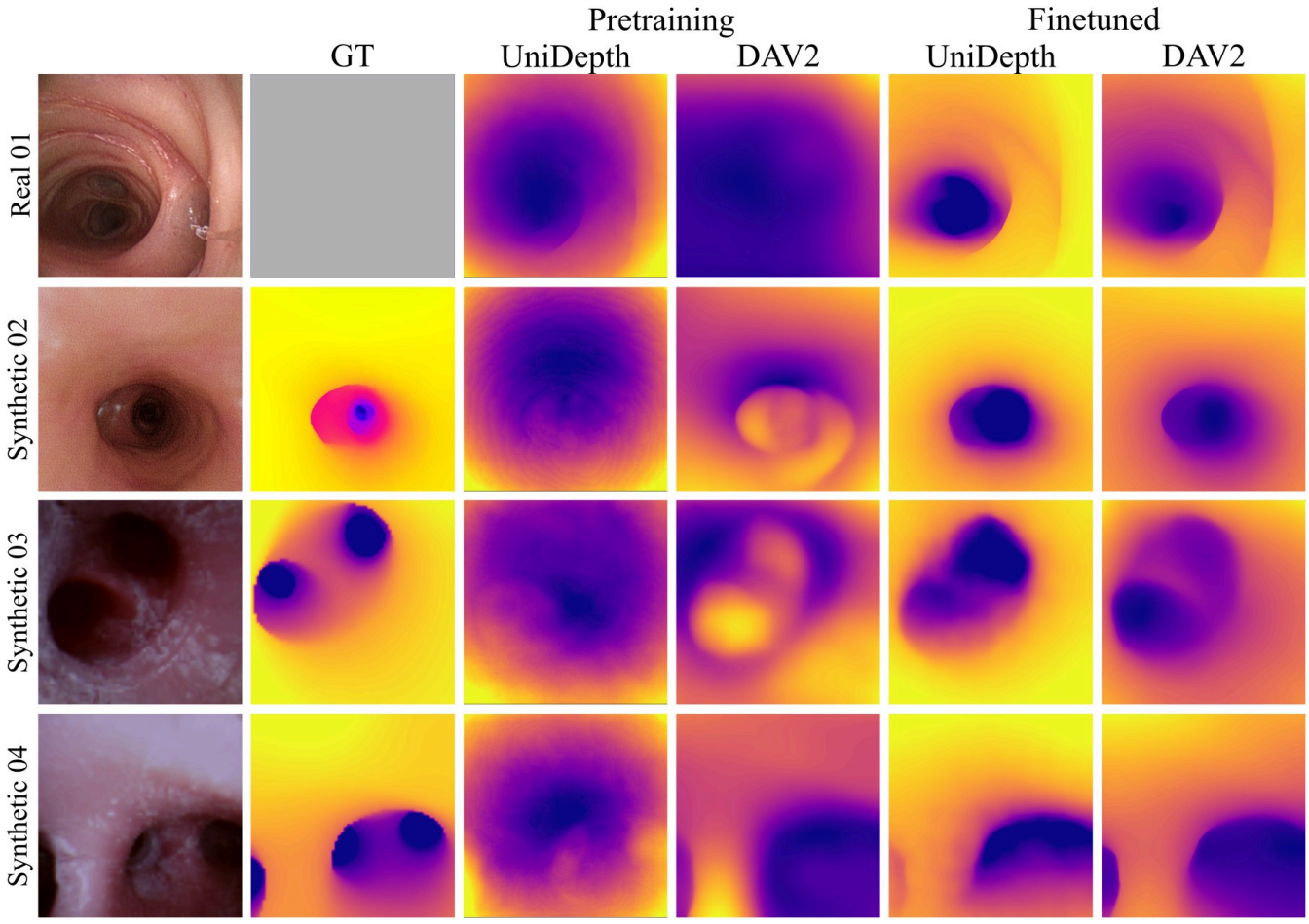

Fig. 7: **Monocular depth estimation examples of pre-trained models and fine-tuned on ROOM.** We show examples on a phantom-based dataset with ground truth [36] as well as real images. Please note that the real image does not have depth ground truth available.

on top of the PyBullet simulator to provide a physically-accurate environment for data collection, it might not fully reflect the contact and deformable dynamics of real bronchia. Enabling support for tissue deformation modelling as well as physiological dynamics such as respiratory motion might also provide realism to the synthesised data.

Lastly, the physical simulator can enable future research in closed-loop navigation systems, enabling its use for validating traditional planners, or for developing imitation learning or reinforcement learning-based navigation policies, as proposed by recent works [19].

VI. Conclusions

We introduced ROOM, a physics-based simulation framework that addresses the critical data scarcity challenge in bronchoscopy robotics. By integrating patient-specific anatomical reconstruction, continuum robot physics, and photorealistic rendering at medically relevant scales, ROOM enables generation of diverse training datasets that capture the complexity of clinical procedures. Our evaluation in established tasks such as multi-view pose estimation and monocular depth estimation reveals that the bronchoscopy domain presents significant challenges for existing methods. However, we showed that the synthetic data generated by ROOM can provide avenues for fine-tuning them and improve performance in real settings.

The ROOM framework will be made available for the community. We expect that its modular architecture will enable researchers to test new CT scans, substitute components, or swap rendering engines, physics simulators, or robot models, opening new avenues for medical robotics research.

References

[1] P. E. Dupont, N. Simaan, H. Choset, and C. D. Rucker, "Continuum Robots for Medical Interventions," *Proceedings of the IEEE*, 2022.

[2] K. Greff, F. Belletti, L. Beyer, C. Doersch, Y. Du, D. Duckworth, D. J. Fleet, D. Gnanapragasam, F. Golemo, C. Herrmann, T. Kipf, A. Kundu, D. Lagun, I. Laradji, H.-T. D. Liu, H. Meyer, Y. Miao, D. Nowrouzezahrai, C. Oztireli, E. Pot, N. Radwan, D. Rebain, S. Sabour, M. S. M. Sajjadi, M. Sela, V. Sitzmann, A. Stone, D. Sun, S. Vora, Z. Wang, T. Wu, K. M. Yi, F. Zhong, and A. Tagliasacchi, "Kubric: A Scalable Dataset Generator," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022.

[3] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "TartanAir: A Dataset to Push the Limits of Visual SLAM," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2020.

[4] A. Rau, S. Bano, Y. Jin, P. Azagra, J. Morlana, R. Kader, E. Sanderson, B. J. Matuszewski, J. Y. Lee, D.-J. Lee, E. Posner, N. Frank, V. Elangovan, S. Raviteja, Z. Li, J. Liu, S. Lalithkumar, M. Islam, H. Ren, L. B. Lovat, J. M. Montiel, and D. Stoyanov, "SimCol3D – 3D reconstruction during colonoscopy challenge," *Medical Image Analysis*, 2024.

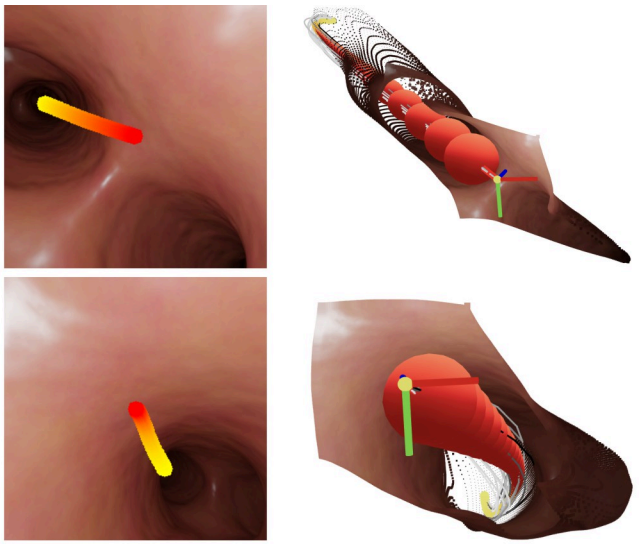

Fig. 8: **Vision-based navigation examples.** We demonstrate qualitative results of the relative monocular depth predictions (scaled with ground-truth scale), as input for a sampling-based local planner. *Left:* projection of the collision-free path. *Right:* 3D visualisation of the path, with the spheres indicating the collision-bodies used by the planner.

[5] Q. Yu, M. Moghani, K. Dharmarajan, V. Schorp, W. C.-H. Panitch, J. Liu, K. Hari, H. Huang, M. Mittal, K. Goldberg, and A. Garg, "ORBIT-Surgical: An Open-Simulation Framework for Learning Surgical Augmented Dexterity," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2024.

[6] E. Coevoet, T. Morales-Bieze, F. Largillière, Z. Zhang, M. Thieffry, M. Sanz-Lopez, B. Carrez, D. Marchal, O. Goury, J. Dequidt, and C. Duriez, "Software toolkit for modeling, simulation and control of soft robots," *Advanced Robotics*, 2017.

[7] S. M. H. Sadati, S. E. Naghibi, A. Shiva, B. Michael, L. Renson, M. Howard, C. D. Rucker, K. Althoefer, T. Nanayakkara, S. Zschaler, C. Bergeles, H. Hauser, and I. D. Walker, "TMTDyn: A Matlab package for modeling and control of hybrid rigid–continuum robots based on discretized lumped systems and reduced-order models," *Intl. J. of Robot. Res.*, 2021.

[8] L. M. Sutherland, P. W. Middleton, A. Russell, M. Wijenayake, N. Maddern, and G. J. Maddern, "Surgical Simulation: A Systematic Review," *Annals of Surgery*, 2006.

[9] C. H. Park, M. J. Ryou, and C. C. Thompson, "Simulation in Endoscopy: Practical Educational Strategies to Improve Learning," *World Journal of Gastroenterology*, 2019.

[10] Y. Liu, C. Li, C. Yang, and Y. Yuan, "EndoGaussian: Real-time Gaussian Splatting for Dynamic Endoscopic Scene Reconstruction," *arXiv preprint arXiv:2401.12561*, 2024.

[11] C. Li, H. Liu, Y. Liu, B. Y. Feng, W. Li, X. Liu, Z. Chen, J. Shao, and Y. Yuan, "Endora: Video Generation Models as Endoscopy Simulators," in *Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2024.

[12] L. Ros-Freixedes, A. Gao, N. Liu, M. Shen, and G.-Z. Yang, "Design optimization of a contact-aided continuum robot for endobronchial interventions based on anatomical constraints," *International Journal of Computer Assisted Radiology and Surgery*, 2019.

[13] Q. Tian, Z. Chen, H. Liao, X. Huang, B. Yang, L. Li, and H. Liu, "PANS: Probabilistic Airway Navigation System for Real-time Robust Bronchoscope Localization," in *Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, 2024.

[14] J. Deng, P. Li, K. Dhaliwal, C. X. Lu, and M. Khadem, "Feature-based Visual Odometry for Bronchoscopy: A Dataset and Benchmark," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[15] J. Borrego-Carazo, C. Sánchez, D. Castells-Rufas, J. Carrabina, and D. Gil, "BronchoPose: an analysis of data and model configuration for vision-based bronchoscopy pose estimation," *Computer Methods and Programs in Biomedicine*, 2023.

[16] V. Vu et al., "BM-BronchoLC: A rich bronchoscopy dataset for anatomical landmarks and lung cancer lesion recognition," *Scientific Data*, 2024.

[17] R. Hao et al., "UAAL Dataset: Upper Airway Anatomical Landmark Dataset for Automated Bronchoscopy and Intubation," *Scientific Data*, 2024.

[18] J. Zhang, L. Liu, P. Xiang, Q. Fang, X. Nie, H. Ma, J. Hu, R. Xiong, Y. Wang, and H. Lu, "AI Co-Pilot Bronchoscope Robot," *Nature Communications*, 2024.

[19] J. Zhao, H. Chen, Q. Tian, J. Chen, B. Yang, and H. Liu, "BronchoCopilot: Towards Autonomous Robotic Bronchoscopy via Multimodal Reinforcement Learning," *arXiv preprint arXiv:2403.01483*, 2024.

[20] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A Self-configuring Method for Deep Learning-based Biomedical Image Segmentation," *Nature Methods*, 2021.

[21] A. Tagliasacchi, T. Delame, M. Spagnuolo, N. Amenta, and A. Telea, "3D skeletons: A state-of-the-art report," *Computer Graphics Forum*, 2016.

[22] J. L. Schönberger and J.-M. Frahm, "Structure-from-Motion Revisited," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4104–4113.

[23] "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM."

[24] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUSt3R: Geometric 3D Vision Made Easy," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[25] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual Geometry Grounded Transformer," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025.

[26] P. Azagra *et al.*, "Endomapper dataset of complete calibrated endoscopy procedures," *Scientific Data*, 2023.

[27] R. Elvira, J. D. Tardós, and J. M. Montiel, "CudaSIFT-SLAM: multiple-map visual SLAM for full procedure mapping in real human endoscopy," *arXiv preprint arXiv:2405.16932*, 2024.

[28] Z. Li *et al.*, "Pose estimation via structure-depth information from monocular endoscopy images sequence," *Optica Publishing Group*, 2024.

[29] J. Klapper, S. Raja, N. Ninan, and S. Shofer, "Bronchoscopy," TSRA Primer in Cardiothoracic Surgery, The American Association for Thoracic Surgery, 2024.

[30] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, "Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

[31] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," in *Advances in Neural Information Processing Systems*, 2024.

[32] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "UniDepth: Universal Monocular Metric Depth Estimation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024.

[33] B. Cui, M. Islam, L. Bai, A. Wang, and H. Ren, "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*. Springer, 2024, pp. 208–218.

[34] Q. Tian, Z. Chen, H. Liao, X. Huang, L. Li, S. Ourselin, and H. Liu, "EndoOmni: Zero-shot cross-dataset depth estimation in endoscopy by robust self-learning from noisy labels," *arXiv preprint arXiv:2409.05442*, 2024.

[35] F. X. Zhang, E. Mackute, M. Kasaei, K. Dhaliwal, R. Thomson, and M. Khadem, "BREA-Depth: Bronchoscopy Realistic Airway-geometric Depth Estimation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2025*, 2025.

[36] M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, and S. Koto, "Deep monocular 3D reconstruction for assisted navigation in bronchoscopy," *International journal of computer assisted radiology and surgery*, vol. 12, pp. 1089–1099, 2017.

[37] J. Jankowski, L. Brudermüller, N. Hawes, and S. Calinon, "VP-STO: Via-point-based Stochastic Trajectory Optimization for Reactive Robot Behavior," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2023, pp. 10 125–10 131.