Feature-based Visual Odometry for Bronchoscopy: A Dataset and Benchmark

Jianning Deng, Peize Li, Kevin Dhaliwal, Chris Xiaoxuan Lu*, and Mohsen Khadem*

***Abstract*— Bronchoscopy is a medical procedure that involves the insertion of a flexible tube with a camera into the airways to survey, diagnose and treat lung diseases. Due to the complex branching anatomical structure of the bronchial tree and the similarity of the inner surfaces of the segmental airways, navigation systems are now being routinely used to guide the operator during procedures to access the lung periphery. Current navigation systems rely on sensor-integrated bronchoscopes to track the position of the bronchoscope in real-time. This approach has limitations, including increased cost and limited use in non-specialized settings. To address this issue, researchers have proposed visual odometry algorithms to track the bronchoscope camera without the need for external sensors. However, due to the lack of publicly available datasets, limited progress is made. To this end, we have developed a database of bronchoscopy videos in a phantom lung model and ex-vivo human lungs. The dataset contains 34 video sequences with over 23,000 frames with odometry ground truth data collected using electromagnetic tracking sensors. With our dataset, we empower the robotics and machine learning community to advance the field. We share our insights on challenges in endoscopic visual odometry. Furthermore, we provide benchmark results for this dataset. State-of-the-art feature extraction algorithms including SIFT, ORB, Superpoint, Shi-Tomasi, and LoFTR are tested on this dataset. The benchmark results demonstrate that the LoFTR algorithm outperforms other approaches, but still has significant errors in the presence of rapid movements and occlusions.**

I. Introduction

Lung cancer is the leading cause of cancer-related deaths, with an overall 5-year survival rate of 17% after diagnoses [1]. This has led to low-dose screening with CT, which has shown benefit in identifying suspected tumours in at-risk patients at an early stage. Once identified, these patients undergo either surveillance CT scans or interventional procedures such as bronchoscopy for characterising and diagnosing the suspected tumour. Bronchoscopy is a procedure to directly examine the airways in the lungs using a small tube with a camera. In addition to cancer biopsy, bronchoscopy can be used for inspection of possible infection or injury in the lung, and local delivery of various treatment modalities such as therapeutic ablation.

Advanced imaging techniques such as computed tomography (CT) scans are employed in the development of navigation systems to assist physicians in navigating within the lungs during bronchoscopy procedures. These systems create a 3D map of the patient's lungs that is used to guide the bronchoscope to the desired location. The navigation systems utilize specialized sensor-integrated bronchoscopes, such as Electro Magnetic Tracker sensors (EMT) or Fiber Bragg Grating shape sensors (FBG), to track the movement and position of the bronchoscope in real time. Examples of navigation platforms that utilize EMT sensors include the Monarch robotic bronchoscopy platform (Auris Health, Inc., USA) [2], superDimension (Medtronic, USA) [3], and SPiN Navigation Systems (Olympus, USA) [4], while Ion (Intuitive Surgical, USA) [5] employs FBG technology for tracking the bronchoscope's pose and shape. However, these specialized sensor-integrated bronchoscopes are significantly more expensive than disposable bronchoscopes and require additional equipment, such as a magnetic field generator (for EMT-based tracking) or FBG fires and an optical interrogator (for FBG-based tracking). Additionally, EMT-based navigation systems have certain limitations [6]. For example, EMT sensors can be influenced by ferromagnetic tools in the operation theatre, which restricts their use in general surgical suites or critical care units.

In this work, we investigate the application of state-of-the-

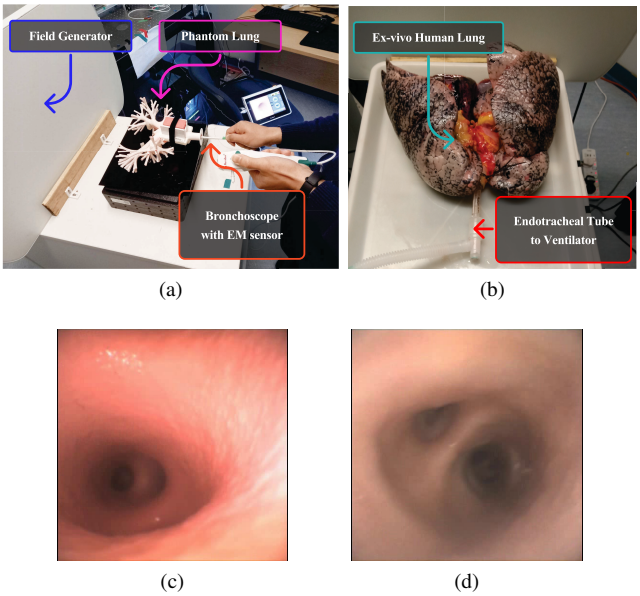


Fig. 1: Experimental Setup used for data collection in (a) phantom lung, (b) mechanically ventilated ex-vivo human lung. Example frames in (c) phantom lung and (d) ex-vivo human lung.

This work was supported by the Medical Research Council [MR/T023252/1].

J. Deng (corresponding author: jianning.deng@ed.ac.uk), P. Li, C. Lu, and M. Khadem are with the School of Informatics, University of Edinburgh, UK. M. Khadem and K. Dhaliwal are with the Translational Healthcare Technologies Group in Centre for Inflammation Research, The Queen's Medical Research Institute, University of Edinburgh, UK. * denotes equal senior authorship.

art visual odometry algorithms for bronchoscope tracking. Visual odometry is a term used in robotics to describe a method of estimating an agent's position and movement using feedback from a camera placed on the agent. The proposed scheme does not rely on additional sensors and only uses the feedback of the endoscopic camera of a manual bronchoscope. Visual odometry algorithms have been previously proposed for other endoscopic applications, such as Colonoscopy [7], [8] or Laparoscopic surgery [9]. However, it has not been explored for bronchoscopy. In this paper, we present the first public dataset captured from several bronchoscopy procedures performed by expert pulmonologists on a phantom lung model and a mechanically ventilated ex-vivo human lung. The dataset encompasses more than 17000 images from 30 video sequences with full pose information for the tip of the bronchoscope. Moreover, to benchmark visual odometry in bronchoscopy, we implemented several feature extraction algorithms including SIFT, ORB, Superpoint, Shi-Tomasi, and a novel detector-less machine learning algorithm, namely, LoFTR. We present results to serve as a benchmark for future research.

II. Related Work

A. Endoscopy dataset

In recent years, considerable progress has been made in robot perception tasks such as classification, segmentation, and pose estimation. This advancement can be attributed to the availability of large-scale datasets that provide the fundamental basis for developing these tasks. However, due to the unique application environment of medical robotics, most existing datasets are unsuitable for this field. Visual data obtained from endoscopy procedures exhibit rare similarities to standard indoor/outdoor environments, and the collected data may differ significantly among organs due to their distinct biological appearances. Consequently, the research in medical robotics perception requires a large variety of datasets to cover different tasks for each organ. Despite the release of several datasets to support visual perception tasks in medical robotics [8], [10]–[15], more datasets are needed to support the research community. Details of these datasets are listed in Tab. I. For pose estimation, both [15] and [8] provide data for visual pose estimation and mapping. [8] is the only publicly available dataset including different organs, [15] can only be accessed by request for colonoscopy pose estimation and mapping. Compared to other organs for existing datasets, the topological geometry for bronchi is much more complex in the lung as they divide hierarchically into branches. Additionally, the walls in the airway are much smoother, with less distinguishable visual features making pose estimation for the bronchial division highly challenging.

B. Feature-based visual odometry

The feature-based approach is one of the popular methods used for visual odometry. This approach focuses on detecting and matching salient pixel points in the images. Once the keypoint matches are set, rotation and translation between frames can be estimated by solving the essential matrix. The feature-based approach is robust to distortions and illumination changes but highly relies on a suitable detector and descriptor for different scenarios. Prominent traditional feature extractors include the ORB [16] features used by ORB-SLAM [17], Shi-Tomasi [18] used by VINS-mono [19], and BRISK used by OKVIS [20]. Together with SIFT [21] and SURF [22] features which also show comparable performance [23], these hand-crafted features have shown great success for robot pose estimation in indoor/outdoor environments.

In addition to model-based approaches, learning-based features have also gained significant attention from the community and have started to outperform the hand-crafted features in terms of accuracy and robustness [24]–[27]. Notably, SuperPoint [24] achieves training in a self-supervised manner using synthetic data and image warping, with a subsequent matching network proposed later [28]. Another notable model, LoFTR [25], replaces the traditional detection-description-matching pipeline with a coarse-to-fine direct correspondence structure to achieve fine pixel-level matching.

C. Inner-body Visual Odometry

Recently, there has been a growing interest in utilizing feature-based visual odometry or Simultaneous Localisation and Mapping (SLAM) techniques for navigating within the human body, as demonstrated in various studies [29]–[33]. While early works employed model-based feature detectors such as FAST and Shi-Tomasi [7], [29], recent studies have utilized modified ORB-SLAM frameworks for in-vivo scenes [30]–[32], [34]. These studies have reported promising results with the ORB detector in coping with the endoscopy environment, including successfully tracking several hundred frames for bronchoscope videos and reconstructing local 3D maps [31], [34]. However, the performance of learning-based feature detectors for inner body videos still needs to be explored. This study aims to benchmark various feature detectors for monocular visual odometry and explore the potential of learning-based methods for this task.

III. Data Collection

A. Experimental setup

We use an Ambu aScope 4 Broncho (Ambu Ltd., Denmark) bronchoscope to record the video stream for our bronchoscopy dataset. The bronchoscope provides a video stream of 15Hz frame rate and 480x480 resolution. Experienced pulmonologists used the bronchoscope to navigate various regions of the lung. Two lung models were used, namely, a bronchoscopy training model (Koken Co., Japan) and a mechanically ventilated ex-vivo human lung.

Meanwhile, we measure the 6-degree-of-freedom (6DoF) ground truth trajectory using the NDI Aurora electromagnetic (EM) tracking system (NDI, Canada). The EM tracking system consists of a field generator that produces a magnetic field and a sensor unit that detects its 6DoF poses within the magnetic field. The miniaturized EM sensor is rigidly connected to the end of the bronchoscopy probe. The EM

Dataset	Organs	Tasks	Availability
Kvasir [10]	colon	anatomical landmarks, phatological findings	Open Academic
Kvasir Seg [11]	colon	polyb segmentation	Open Academic
Endoscopic Artifact Detection [12]	colon	artifact detection	Open Academic
Nerthus [13]	bowel	bowel preparation	Open Academic
EndoAbs [14]	liver, kidney, spleen	stereo 3D reconstruction	By Request
EndoMapper [8]	colon	vSLAM	By Request
EndoSLAM [15]	colon, stomach, small intestine	vSLAM	Open Access
Ours	**lung**	**visual odometry**	**Open Access**

TABLE I: Summary of existing datasets for medical robotics application.

tracking system can provide a ground truth trajectory for the camera within the organ with a mean accuracy of 0.8 mm.

B. Ethics Statement

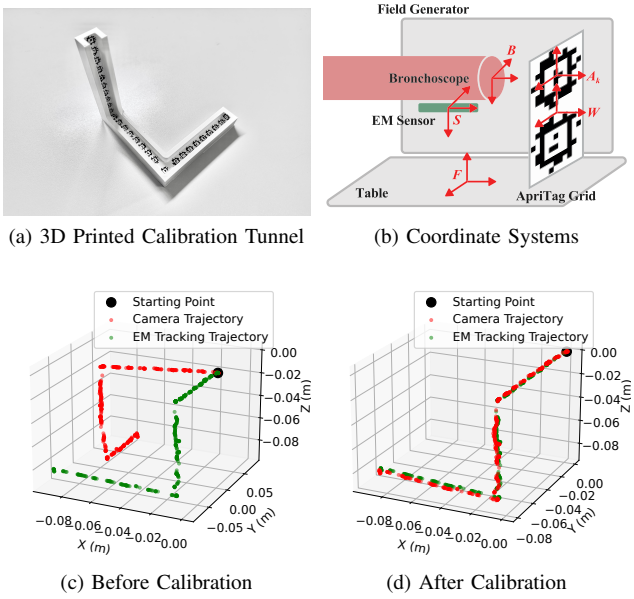
The studies involving human participants were reviewed and approved by London - Central Research Ethics Committee 16-LO-1883. Written informed consent was not provided because this study involves the use of human lungs that are unfit for transplant. Specialist Nurses for Organ Donation (SNODs) approach relatives of potential donors and obtain authorisation for the use of the donor's organs and tissues for transplantation using the standard NHS Authorisation form. Using the same authorisation form, the SNODs also obtain authorisation for the use of the donor's organs and tissues removed but are subsequently found to be unsuitable for transplantation for other purposes (i.e. research studies, education, training). The use of lungs in this study has been approved by an independent ethics committee and NHSBT, the body responsible for organ donation and transplantation across the UK.

C. Calibration

Due to the small diameter of the bronchoscope insertion cord and the EM tracking sensor, the relative translative position between the two is minimal. For this reason, we can equate the translation motion of the camera and the EM sensor without calibrating their relative translation. However, the relative orientation between the camera and the EM sensor will need to be calibrated in order to align the orientation of the camera trajectory and the ground truth trajectory.

During our tests, checkerboard-based methods as seen in [35], [36] are very inaccurate. Due to the shallow depth of field of the bronchoscopy camera and its unadjustable focus distance, the camera focuses only on nearby artefacts while visual features outside a few centimetres can appear blurry. This will limit the camera's movement for the traditional pattern-based calibration methods. Here, we propose a novel calibration method using fiducial markers on a 3d-printed track.

We designed a 3D tunnel with two L shapes stitched together to guide movements in all x-y-z directions, as seen in Fig. 2a. The inside of the tunnel is lined with AprilTag markers [37]. During calibration, the bronchoscopy camera's optical axis maintains approximately perpendicular to the markers, and its movement is guided by the tunnel walls. This way, the camera can move long distances along 3 axes of the space while still being tracked by AprilTag markers.



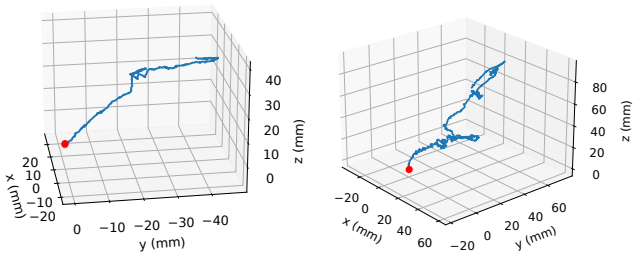
(a) 3D Printed Calibration Tunnel (b) Coordinate Systems

(c) Before Calibration (d) After Calibration

Fig. 2: (a) Calibration tunnel used for registering camera trajectory to EM tracking trajectory. (b) Calibration Setup. EM sensor is rigidly connected to the bronchoscope. April tags placed inside the calibration tunnel are used to estimate camera trajectory independent of the EM sensor. Camera trajectory (c) before and (d) after calibration is compared with EM tracking trajectory.

The calibration setup and coordinate system definitions are shown in Fig. 2b. There are 29 AprilTag markers lined inside the calibration tunnel, each with its own coordinate system, denoted as $(A_k)_{k\in(1,\dots,29)}$. The world coordinate system W is attached to the first tag so that all of the tags have known transformations with W, denoted by $({}^{A_k}\mathbf{T}_W)_{k\in(1,\dots,29)}$. In the meantime, the EM tracking system records trajectories in the Field Generator's coordinate system F. For calibration, we want to calculate the transformation ${}^{B}\mathbf{T}_S$ between the bronchoscope system B and the EM sensor system S.

In a recorded calibration sequence, n frames of 6DoF transformations $({}^{S_i}\mathbf{T}_F)_{i\in(1,\dots,n)}$ are recorded by the EM tracking system, while m frames of images are recorded by the bronchoscope which by using the AprilTag library can we derive the transformations $({}^{B_j}\mathbf{T}_{A_k})_{j\in(1,\dots,m)}$ (k here depends on whichever AprilTag is in view of the bronchoscope in each frame). Due to the higher frame rate of the EM tracking system, points are sampled from the EM tracking sequence based on the closest timestamp so that a one-to-one correspondence between camera and EM tracking can



(a) Phantom lung trajectory (b) Ex-vivo human lung trajectory

Fig. 3: Example of trajectories in our collected data on (a) phantom lung and (b) ex-vivo human lung. The red dot in the figure denotes the starting point. All units in millimetre (mm).

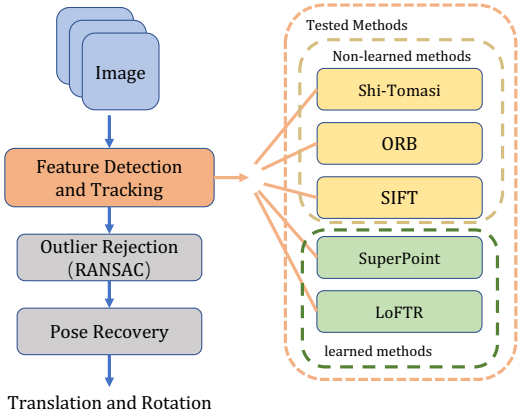


Fig. 4: Overview of the visual odometry pipeline used for benchmarking.

be formed. After sampling, the EM tracking transformations becomes $({}^{S_i}\mathbf{T}_F)_{i\in(1,\dots,m)}$. With these transformations, we can now align the bronchoscope trajectory and EM tracking trajectory using

$$ {}^{B_j}\mathbf{T}_{S_j}\,{}^{S_j}\mathbf{T}_F = {}^{B_j}\mathbf{T}_{A_k}\,{}^{A_k}\mathbf{T}_W \tag{1}$$

where ${}^{B_j}\mathbf{T}_{S_j}$ is the calibration matrix from the EM sensor to the bronchoscope in the j-th frame. The calibration matrix ${}^{B_j}\mathbf{T}_{S_j}$ is derived by aligning the camera trajectory and the sampled EM tracking trajectory using the Kabsch–Umeyama algorithm [38]. Camera calibration results are shown in Fig. 2.

D. Trajectories

To evaluate the performance of various algorithms on bronchoscopy videos, we collected both real lung and phantom lung data for evaluation. Specifically, we record 34 sequences in total for both phantom lung and ex-vivo lung, which contain about 23,000 frames for evaluation purposes. The recorded phantom lung sequences are classified into two difficulty levels, easy and difficult, to allow researchers to evaluate their customized algorithms. While all sequences are captured using a handheld bronchoscope, the easy sequences are recorded with minimal camera movement to ensure stability. In contrast, the difficult sequences aimed to replicate the dynamic movements encountered during actual bronchoscopy. Consequently, changes in illumination, motion blur, and hand-held vibrations are less significant in the easy sequences than in the difficult sequences. Furthermore, to enhance the understanding of the benchmarked methods on real-world data, we also included benchmark results with sequences on real lungs captured by professional medical staff. Example trajectories and images from the collected data are shown in Fig. 3.

IV. Benchmarks

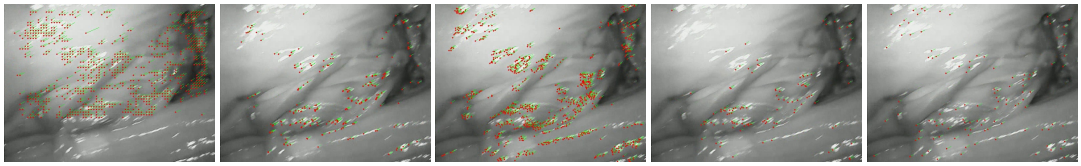
A. Monocular Visual Odometry

The initial step in feature-based monocular visual odometry involves detecting keypoints in the input image at timestep t using a keypoint detector, which returns the coordinates of the keypoints and their corresponding feature vectors. Following this, keypoints are matched between consecutive frames based on their descriptors. Following the visual odometry algorithm proposed in [17], we employ the KNN (k-nearest neighbour) brute force matcher with an L2 distance metric in this work. The matcher yields a set of matching pairs between keypoints in consecutive frames. We estimate the essential matrix using the RANSAC [39] algorithm, which eliminates outliers and requires a minimum of five matching pairs. Once the essential matrix is obtained, we decompose it using SVD (singular value decomposition) to obtain the rotation and translation components of the pose transformation. It is important to note that the scale of the estimated transformation is unknown in the monocular setting. We, therefore, employ the ground truth transformation to recover the scale for the estimated results. The scale is computed by dividing the norm of the translation vectors for the ground truth pose between consecutive frames with the estimated pose transformation. This scale factor is then applied to the estimated translation vector. For the experiments conducted in this paper, we exclusively vary the keypoint detectors while keeping the remaining framework unchanged. In this way, we ensure a fair comparison between the detectors, except for the LoFTR detector, which does not necessarily need a matcher for keypoint matching. The pipeline is illustrated in Fig. 4.

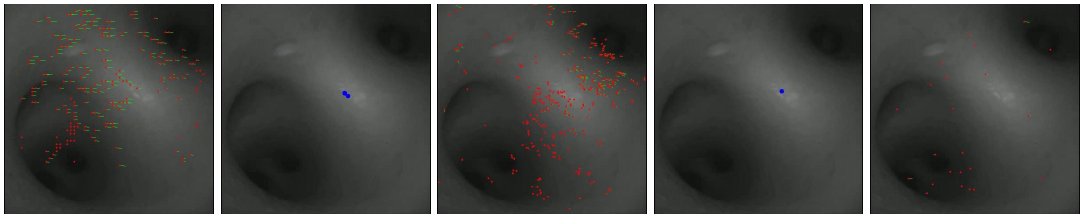
B. Methods and datasets

In order to carry out a comprehensive benchmark evaluation of keypoint detector methods, we have carefully selected a range of approaches that each represent a distinct technique.

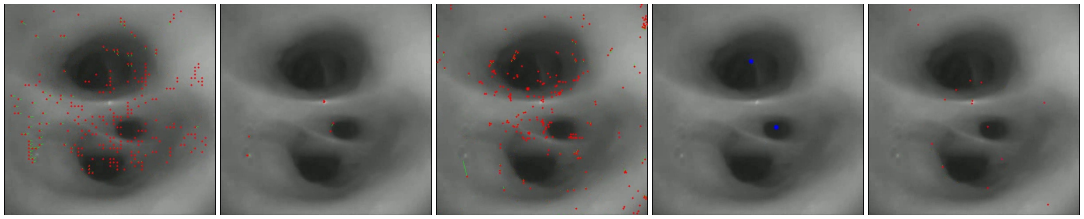
Among non-learned detectors, Shi-Tomasi identifies corners in an image using the minimum eigenvalue of the local gradient matrix. In this benchmark, we follow a similar implementation of Shi-Tomasi from [7]. Conversely, ORB first identifies FAST corner features before describing each feature using a BRIEF descriptor. While ORB and Shi-Tomasi have demonstrated plausible performance for inner-



(a) Colon



(b) Model Lung



(c) Real Lung

Fig. 5: A visualization is provided to showcase the feature tracking performance of various methods across different datasets. The methods are arranged from right to left: LoFTR, ORB, Shi-Tomasi, SIFT, and SuperPoint. The tracked inliers in the current frame are represented as red dots, while the motion from the previous frame to the current frame is depicted as green lines. Additionally, blue dots represent the detected points, which are only shown in frames where tracking failed.

body visual odometry [7], [34], the performance of another popular method, SIFT, remains to be discovered. SIFT detects keypoints in an image at varying scales and orientations, and then describes each keypoint using a feature descriptor invariant to changes in scale, rotation, and illumination. We select the three mentioned representative detectors for the benchmark.

Additionally, we have included SuperPoint, a learning-based feature detector that efficiently detects and describes keypoints in an image. SuperPoint [24], which is trained in a self-supervised manner, uses a CNN backbone to predict the coordinates and descriptors of keypoints. Lastly, we have chosen LoFTR [25] for the direct image registration algorithm. LoFTR employs a convolutional neural network backbone to extract features and learns to match local image patches directly by utilizing transformer layers, thus eliminating the requirement for identifying salient features. The algorithm generates image matching in a coarse-to-fine manner. It first identifies the matching pixels between two images in the $\frac{1}{8}$ resolution. Once the coarse matching pairs are established, the corresponding local feature maps are cropped in the full-scale feature map. Subsequently, spatial offsets in the current images' x and y directions are estimated to obtain precise matching positions with respect to the patch centre of the previous image.

In order to obtain a more profound understanding and a thorough evaluation of the aforementioned methods, we have also conducted experiments on the publicly available EndoSLAM dataset, which comprises real-world data for the colon and stomach and synthetic colonoscopy data. To ensure impartiality, we employed pre-trained models for the learned methods and kept all tunable parameters for non-learned methods at default values. Fig. 4 shows the overall benchmarking pipeline and various odometry algorithms used in this work.

C. Evaluation Metrics

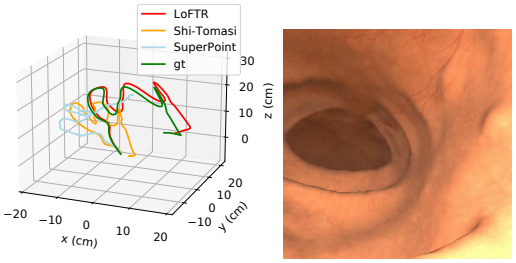
Relative pose error (RPE) we use the relative pose error (RPE) as the primary metric for quantitative evaluation. The RPE measures the pose estimation accuracy over a fixed time interval. Given estimated poses from a video sequence as $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n \in SE(3)$ and ground truth poses of the same sequence $\mathbf{Q}_1, \mathbf{Q}_2, \ldots, \mathbf{Q}_n \in SE(3)$ with a subscript to denote the corresponding frame number, the RPE E_t at time step t can be written as follows:

$$\mathbf{E}_t := (\mathbf{Q}_t^{-1}\mathbf{Q}_{t+1})^{-1} \cdot (\mathbf{P}_t^{-1}\mathbf{P}_{t+1}) \tag{2}$$

Here, we set the time interval as 1 and use RPE for pose transformation between each frame to avoid accumulative drift. Specifically, we compute the root-mean-square error (RMSE) for all time steps for the translation component as:

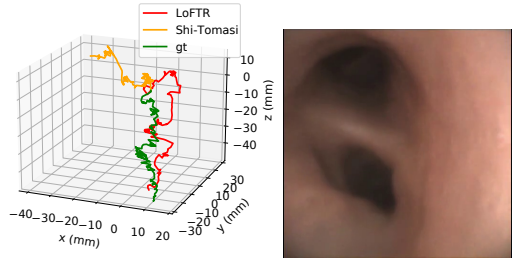
$$RMSE(\mathbf{E}_{trans}) := \sqrt{\frac{1}{N}\sum_{t=1}^{N} ||trans(\mathbf{E}_t)||_2} \tag{3}$$

where the $trans(\mathbf{E}_i)$ denotes the tranlation part of the relative pose error. As for the rotation component of RPE,



(a) Trajectory (b) Screenshot

Fig. 6: Trajectory estimated by different methods on synthetic colonoscopy data. The trajectory here is shown in meters (m).



(a) Trajectory (b) Screenshot

Fig. 7: Trajectory estimated by different methods on real lung seq 3. The trajectory here is shown in millimeter (mm).

we report it in degree with mean and standard deviation (std). **Absolute trajectory error (ATE)** Another evaluation metric used in the work is the absolute trajectory error (ATE). ATE evaluates the global consistency of the estimated trajectory. Note that we are in the monocular setting for most of the in-vivo scenarios, we need to address the scale-missing problem to calculate the ATE. To handle this, we rescale the magnitude of the estimated translation at each step based on the ground truth. We first calculate the Euclidean distance of the ground truth position at consecutive timestamps, denoted as s_{gt}. Second, we calculate the magnitude of the estimated translation, denoted as s_{pred}. Then we apply the scaling factor $\frac{s_{gt}}{s_{pred}}$ on the estimated translation. By doing so, we can retrieve the absolute scale given ground truth as reference and compute the ATE between two trajectories. The ATE at time step t is formulated as follows:

$$\mathbf{F}_t := \mathbf{Q}_t^{-1}\mathbf{S}\mathbf{P}_t \tag{4}$$

where $\mathbf{S}$ is the rigid-body transformation to align two trajectories. Since we already retrieve the scale problem mentioned above, the transformation $\mathbf{S}$ is the same as the extrinsic parameters obtained in Sec. III-C.

Similarly to RPE, we report the root mean square error (RMSE) of the ATE over all time steps as:

$$RMSE(\mathbf{F}) := \sqrt{\frac{1}{N}\sum_{t=1}^{n} ||trans(\mathbf{F}_t)||^2} \tag{5}$$

D. Results

In order to gain deeper insight into the performance of feature detectors, we propose a two-fold evaluation containing measures of both robustness and accuracy. Specifically, robustness is evaluated by determining the percentage of successfully tracked frames within a given sequence. This metric indicates the detector's ability to reliably identify and track key points across consecutive frames, thereby facilitating accurate pose estimation. Conversely, accuracy is assessed by examining whether the detector can consistently generate comparable descriptors for a given point across consecutive frames, even when confronted with changes in illumination, occlusion, viewpoint, and other potential sources of variation.

Robustness The Tab. II reveals that the Shi-Tomasi and LoFTR methods are notable for their robustness across various scenarios, having kept all frames tracked across different sequences. Additionally, the SuperPoint detector displays satisfactory robustness within the EndoSLAM dataset, having tracked the entire synthetic colon sequence. Conversely, ORB and SIFT achieve high success rates in the EndoSLAM dataset, except for SIFT in the synthetic colon sequence, yet they fall short in processing the complete sequences. This outcome implies that these detectors are less reliable when confronted with dynamic and feature-scarce scenarios within the inner-body environment. In the bronchoscopy dataset, the success rates for ORB, SIFT, and SuperPoint drop significantly, indicating the difficulty of extracting distinct features from bronchoscopy videos. We believe this is caused by the smooth surface and airway of the bronchi compared to the texture-rich surface of the stomach and colon. Nevertheless, the Shi-Tomasi and LoFTR maintain 100% success rates in the bronchoscopy, further proving their robustness in different in-vivo scenarios. Qualitative results for the feature tracking of different methods are shown in Fig. 5. Based on the visualization shown in Fig. 5, it is evident that LoFTR and Shi-Tomasi have a higher number of detected key points than the other methods, confirming the robustness of these two methods. In contrast, blue dots occur in several scenarios for SIFT, indicating that it fails to detect enough reliable key points for pose recovery.

Accuracy The evaluation of the accuracy for different methods is presented in Tab. III. It is important to note that results are only reported when the method successfully tracks its pose for the entire sequence to ensure fair comparisons. Based on the presented results, we observe that synthetic colonoscopy is the least challenging sequence for visual odometry. Furthermore, LoFTR demonstrates the lowest ATE and RPE in most sequences, highlighting its superior performance across various organs. However, errors for the colon and stomach in the EndoSLAM dataset are exceptionally high. We attribute such error to the frame skipping in the EndoSLAM dataset, which implies that the released videos consist of non-consecutive frames selected at inconsistent intervals. The rationale for such frame selection is unclear, but it increases the differences between frames, making them

Dataset	Organ	Methods				
		ORB [16]	SIFT [21]	Shi-Tomasi [7]	SuperPoint [24]	LoFTR [25]
EndoSLAM [15]	Colon	99.79%	98.64%	**100%**	98.11%	**100%**
	stomach	99.81%	96.71%	**100%**	97.77%	**100%**
	Synthetic Colon	99.8%	59.67%	**100%**	**100%**	**100%**
Ours	Model Lung Easy	21.75%	0.75%	**100%**	91.50%	**100%**
	Model Lung Difficult	25.00%	6.78%	**100%**	75.57%	**100%**
	Real Lung Seq 1	22.97%	14.86%	**100%**	69.23%	**100%**
	Real Lung Seq 2	19.23%	5.77%	**100%**	69.23%	**100%**
	Real Lung Seq 3	0%	0%	**100%**	55.22%	**100%**

TABLE II: Percentage of successfully tracked frames in the sequence.

Dataset	Organ	Methods								
		Shi-Tomasi [7]			SuperPoint [24]			LoFTR [25]		
		ATE	RPE_{Trans}	RPE_{Rot}	ATE	RPE_{Trans}	RPE_{Rot}	ATE	RPE_{Trans}	RPE_{Rot}
EndoSLAM [15]	Colon	167.78	2.44	25.23±61.49	-	-	-	**159.66**	**2.15**	**18.16±52.89**
	Stomach	105.86	1.66	22.30±58.23	-	-	-	**77.94**	**1.65**	**15.15±48.2**
	Synthetic Colon	114.06	0.47	0.84±1.23	170.63	0.62	1.31±1.43	**24.03**	**0.13**	**0.24±0.18**
Ours	Model Lung Easy 1	**18.94**	**0.44**	**2.06±6.19**	-	-	-	20.46	0.45	2.14±12.68
	Model Lung Difficult 1	118.12	1.71	8.78±18.56	-	-	-	**97.64**	**1.70**	**4.82±10.47**
	Real Lung Seq 1	14.64	0.64	4.48±16.01	-	-	-	**5.16**	**0.56**	**2.35±2.38**
	Real Lung Seq 2	**5.34**	**0.93**	5.17±8.37	-	-	-	11.72	0.95	**3.46±3.38**
	Real Lung Seq 3	49.77	2.46	8.44±10.91	-	-	-	**15.24**	**2.29**	**4.47±3.04**

TABLE III: ATE and RPE for different methods on different datasets, where *Trans.* denotes translation part and *Rot.* denotes rotation part. The RMSE of translation part is reported in millimetres (mm) and the rotation part is reported in degree (°) with mean and standard deviation.

difficult to track. Moreover, Tab. III suggest that videos captured from lung models are particularly challenging to track. This is attributed to the fact that phantom data are collected with larger trajectory lengths, leading to larger ATE. Besides, due to the limited flexibility of the probe, we need to do more rotation along the camera z-axis to ensure stable movements, making it more difficult for rotation estimation. For qualitative evaluation, the synthetic colon trajectory is drawn in Fig. 6 for visualization, since three of the tested methods managed to process the entire sequence. With regard to our dataset, the trajectories generated by LoFTR and Shi-Tomasi for real lung sequence 3 are presented in Figure 7.

Comparing the results of the model-based visual feature detectors, it is apparent that Shi-Tomasi outperforms SIFT and ORB in terms of robustness. We credit Shi-Tomasi's simple corner feature detection method for this superior performance, which relies on detecting the minimum eigenvalue of the local gradient matrix. This technique makes Shi-Tomasi more suitable for feature-sparse and low-contrast in-vivo scenes than ORB and SIFT. However, the Shi-Tomasi detector fails to yield strong descriptors for matching, leading to less accurate visual odometry. In contrast, LoFTR utilizes deep neural networks for feature extraction and attention mechanisms to match images. This feature makes LoFTR particularly excel at visual odometry in low-contrast bronchoscopy scenarios, where patch-to-patch matches can be generated without regard to the saliency of the image patch. Nevertheless, the minimal ATE reported is in the sub-centimetre level, indicating that the existing feature extraction methods are not optimal for bronchoscopy.

V. CONCLUSIONS

Motivated by (i) research advancements in inner-body SLAM, and (ii) a lack of publicly available datasets for design and testing of Odometry algorithms in bronchoscopy, we present the first public dataset of lung bronchoscopy procedure in phantom lung and mechanically ventilated ex-vivo human lung. Furthermore, we benchmarked feature-based odometry algorithms on this data. Several algorithms designed for indoor/outdoor robots including SIFT, ORB, Superpoint, Shi-Tomasi, and LoFTR were tested. The results demonstrated that only Shi-Tomasi and LoFTR are robust enough for inner-body visual odometry, while LoFTR shows superior performance in terms of both robustness and accuracy across most of the evaluated sequences. Nonetheless, the current performance levels of the LoFTR algorithm in bronchoscopy visual odometry within the lung are still far from satisfactory, necessitating future work on more robust Odometry algorithms for endoluminal procedures.

REFERENCES

[1] M. Coleman, D. Forman, and *et. al.*, "Cancer survival in australia, canada, denmark, norway, sweden, and the UK, 1995–2007 (the international cancer benchmarking partnership): an analysis of population-based cancer registry data," *The Lancet*, vol. 377, no. 9760, pp. 127–138, Jan. 2011.

[2] J. R. Rojas-Solano, L. Ugalde-Gamboa, and M. Machuzak, "Robotic bronchoscopy for diagnosis of suspected lung cancer," *Journal of Bronchology and Interventional Pulmonology*, vol. 25, no. 3, pp. 168–175, July 2018.

[3] M. P. Rivera, A. C. Mehta, and M. M. Wahidi, "Establishing the diagnosis of lung cancer," *Chest*, vol. 143, no. 5, pp. e142S–e165S, May 2013.

[4] B. S. Furukawa, N. J. Pastis, N. T. Tanner, A. Chen, and G. A. Silvestri, "Comparing pulmonary nodule location during electromagnetic bronchoscopy with predicted location on the basis of two virtual airway maps at different phases of respiration," *Chest*, vol. 153, no. 1, pp. 181–186, Jan. 2018.

[5] J. Reisenauer, M. J. Simoff, M. A. Pritchett, D. E. Ost, A. Majid, C. Keyes, R. F. Casal, M. S. Parikh, J. Diaz-Mendoza, S. Fernandez-Bussy, and E. E. Folch, "Ion: Technology and techniques for shape-sensing robotic-assisted bronchoscopy," *The Annals of Thoracic Surgery*, vol. 113, no. 1, pp. 308–315, Jan. 2022.

[6] X. Luó, M. Feuerstein, T. Kitasaka, and K. Mori, "Robust bronchoscope motion tracking using sequential monte carlo methods in navigated bronchoscopy: dynamic phantom and patient validation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 7, no. 3, pp. 371–387, July 2011.

[7] J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Tracking monocular camera pose and deformation for slam inside the human body," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5278–5285.

[8] P. Azagra, C. Sostres, Á. Ferrandez, L. Riazuelo, C. Tomasini, O. L. Barbed, J. Morlana, D. Recasens, V. M. Batlle, J. J. Gómez-Rodríguez, *et al.*, "Endomapper dataset of complete calibrated endoscopy procedures," *arXiv preprint arXiv:2204.14240*, 2022.

[9] X. Liu, Z. Li, M. Ishii, G. D. Hager, R. H. Taylor, and M. Unberath, "Sage: Slam with appearance and geometry prior for endoscopy," *arXiv preprint arXiv:2202.09487*, 2022.

[10] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, *et al.*, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 164–169.

[11] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. d. Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 451–462.

[12] S. Ali, F. Zhou, B. Braden, A. Bailey, S. Yang, G. Cheng, P. Zhang, X. Li, M. Kayser, R. D. Soberanis-Mukul, *et al.*, "An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy," *Scientific reports*, vol. 10, no. 1, pp. 1–15, 2020.

[13] K. Pogorelov, K. R. Randel, T. de Lange, S. L. Eskeland, C. Griwodz, D. Johansen, C. Spampinato, M. Taschwer, M. Lux, P. T. Schmidt, *et al.*, "Nerthus: A bowel preparation quality video dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 170–174.

[14] V. Penza, A. S. Ciullo, S. Moccia, L. S. Mattos, and E. De Momi, "Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 5, p. e1926, 2018.

[15] K. Bengisu Ozyoruk, G. Irem Gokceler, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira, H. Sahin, *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos: Endo-sfmlearner," *arXiv e-prints*, pp. arXiv–2006, 2020.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[17] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct. 2015.

[18] J. Shi *et al.*, "Good features to track," in *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 1994, pp. 593–600.

[19] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[20] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[21] P. C. Ng and S. Henikoff, "Sift: Predicting amino acid changes that affect protein function," *Nucleic acids research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[22] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[23] H.-J. Chien, C.-C. Chuang, C.-Y. Chen, and R. Klette, "When to use what feature? sift, surf, orb, or a-kaze features for monocular visual odometry," in *2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2016, pp. 1–6.

[24] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[25] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[26] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 467–483.

[27] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler, "D2-net: A trainable cnn for joint description and detection of local features," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2019, pp. 8092–8101.

[28] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[29] O. G. Grasa, E. Bernal, S. Casado, I. Gil, and J. Montiel, "Visual slam for handheld monocular endoscope," *IEEE transactions on medical imaging*, vol. 33, no. 1, pp. 135–146, 2013.

[30] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Computer methods and programs in biomedicine*, vol. 158, pp. 135–146, 2018.

[31] C. Wang, M. Oda, Y. Hayashi, T. Kitasaka, H. Honma, H. Takabatake, M. Mori, H. Natori, and K. Mori, "Visual slam for bronchoscope tracking and bronchus reconstruction in bronchoscopic navigation," in *Medical Imaging 2019: Image-Guided Procedures, Robotic Interventions, and Modeling*, vol. 10951. SPIE, 2019, pp. 51–57.

[32] N. Mahmoud, T. Collins, A. Hostettler, L. Soler, C. Doignon, and J. M. M. Montiel, "Live tracking and dense reconstruction for handheld monocular endoscopy," *IEEE transactions on medical imaging*, vol. 38, no. 1, pp. 79–89, 2018.

[33] L. Qiu and H. Ren, "Endoscope navigation and 3d reconstruction of oral cavity by visual slam with mitigated data scarcity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2197–2204.

[34] C. Wang, M. Oda, Y. Hayashi, B. Villard, T. Kitasaka, H. Takabatake, M. Mori, H. Honma, H. Natori, and K. Mori, "A visual slam-based bronchoscope tracking scheme for bronchoscopic navigation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, pp. 1619–1630, 2020.

[35] T. Reichl, X. Luo, M. Menzel, H. Hautmann, K. Mori, and N. Navab, "Hybrid electromagnetic and image-based tracking of endoscopes with guaranteed smooth output," *International journal of computer assisted radiology and surgery*, vol. 8, pp. 955–965, 2013.

[36] X. Luo, M. Feuerstein, T. Sugiura, T. Kitasaka, K. Imaizumi, Y. Hasegawa, and K. Mori, "Towards hybrid bronchoscope tracking under respiratory motion: evaluation on a dynamic motion phantom," in *Medical Imaging 2010: Visualization, Image-Guided Procedures, and Modeling*, vol. 7625. SPIE, 2010, pp. 410–420.

[37] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2016.

[38] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, Apr. 1991.

[39] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.