

Geometry-Aware Camera Localization for Bronchoscopy

Lumin Chen
lumin.chen@cair-cas.org.hk
Centre for Artificial Intelligence and
Robotics, Hong Kong Institute of
Science & Innovation, Chinese
Academy of Sciences
Hong Kong, China

Qingyao Tian
tianqingyao2021@ia.ac.cn
Institute of Automation, Chinese
Academy of Sciences
Beijing, China

Jinpeng Li
Haoyu Jiang
jinpeng.li@cair-cas.org.hk
haoyu.jiang@cair-cas.org.hk
CAIR, Hong Kong Institute of Science
& Innovation, Chinese Academy of
Sciences
Hong Kong, China

Huai Liao
Xinyan Huang
liaohuai@mail.sysu.edu.cn
hxinyan@mail.sysu.edu.cn
The First Affiliated Hospital, Sun
Yat-sen University
Guangzhou, China

Hongbin Liu
hongbin.liu@cair-cas.org.hk
Centre for Artificial Intelligence and
Robotics, Hong Kong Institute of
Science & Innovation, Chinese
Academy of Sciences
Hong Kong, China

Dong Yi[✉]
dong.yi@cair-cas.org.hk
Centre for Artificial Intelligence and
Robotics, Hong Kong Institute of
Science & Innovation, Chinese
Academy of Sciences
Hong Kong, China

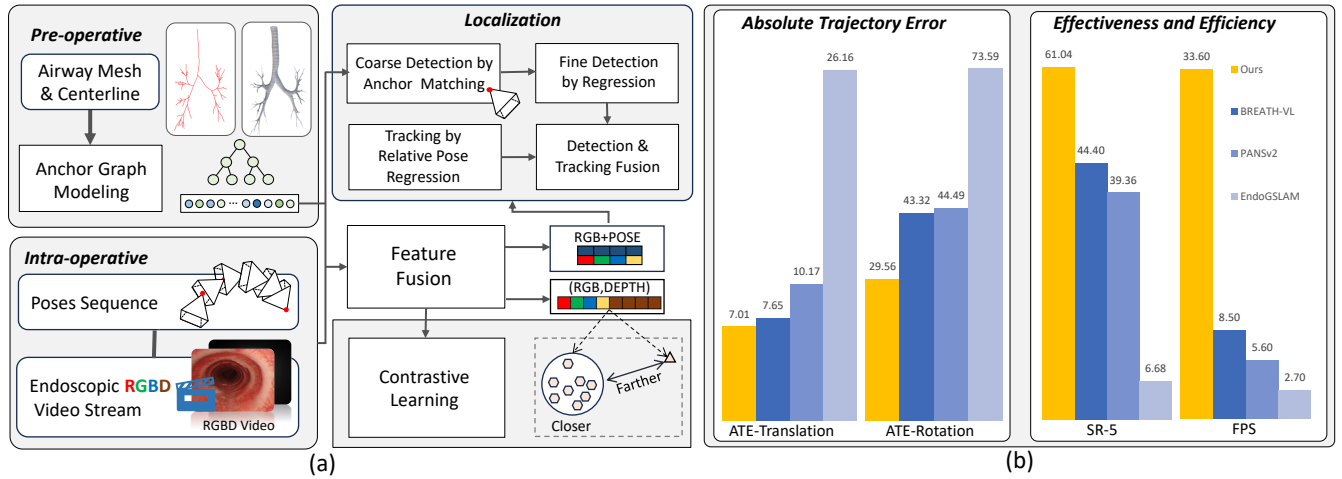


Figure 1: Overview and performance of Geometry-Aware Bronchoscopy Localization framework (GABL). (a) GABL combines preoperative airway geometry with intraoperative observations for anchor localization, temporal tracking, and RGB-depth matching. (b) Comparison with prior bronchoscopy localization methods in accuracy, success rate, and inference speed.

Abstract

Camera localization in bronchoscopy remains a challenging problem due to stringent accuracy requirements, real-time constraints, and limited training data. Compared to natural scenes, the confined anatomical structures demand millimeter-level precision, while intraoperative guidance necessitates low-latency inference. However, existing methods often fail to effectively exploit preoperative geometric priors, limiting their robustness and accuracy. To

address these limitations, we propose a unified geometry-aware bronchoscope localization framework (GABL) that effectively fuses preoperative structural priors with paired intraoperative video to estimate 6-DoF camera poses. Specifically, to address visual ambiguity in complex airways, we propose a graph-guided coarse-to-fine localization scheme that effectively leverages structural priors for precise pose estimation. Furthermore, to mitigate pose jitter and bridge the visual-structural gap, we integrate a Transformer-based tracking model with a novel RGB-depth matching objective, jointly enforcing spatio-temporal and geometric consistency. Extensive experiments demonstrate that our method yields remarkable reductions of 8.37% and 31.76% in translation and rotation errors over the prior state-of-the-art, alongside 4 times inference speedup (33.6 FPS) for robust real-time bronchoscope localization.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil.

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3835259>

CCS Concepts

• **Computing methodologies** → **Vision for robotics**.

Keywords

Camera localization, Geometry-aware localization, Graph neural network

ACM Reference Format:

Lumin Chen, Qingyao Tian, Jinpeng Li, Haoyu Jiang, Huai Liao, Xinyan Huang, Hongbin Liu, and Dong Yi. 2026. Geometry-Aware Camera Localization for Bronchoscopy. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835259>

1 Introduction

Bronchoscopy localization aims to estimate the real-time 6-DoF pose of the endoscope within the airway, enabling accurate navigation and spatial awareness during minimally invasive procedures [1]. This capability is critical for assisting physicians in reaching target regions, improving surgical safety, and enabling advanced applications such as image-guided intervention and augmented reality visualization [38]. However, achieving reliable localization in bronchoscopy is inherently challenging due to the complex and deformable anatomical structures, the prevalence of repetitive and low-texture surfaces, and the extremely limited field of view. These factors make it difficult to establish stable visual correspondences, while the small spatial scale of the airway further imposes stringent requirements on localization accuracy and real-time performance.

Despite significant progress in general camera localization, existing methods are not well suited for bronchoscopic scenarios. Most approaches rely on large-scale annotated datasets and visual features learned from natural images [24, 47], which suffer from severe domain gaps when applied to medical imagery. In addition, techniques designed for outdoor or indoor environments often assume rich textures, wide fields of view, and static scenes, conditions that do not hold in endoscopy. As a result, these methods tend to degrade significantly in the presence of ambiguous structures and dynamic tissue deformation. Consequently, directly applying general-purpose localization frameworks fails to meet the millimeter-level accuracy and robustness required in clinical bronchoscopy [10].

To address these challenges, existing approaches primarily focus on exploiting geometric cues to compensate for the lack of reliable visual features in endoscopic imagery. On the one hand, one line of work reconstructs the anatomical structures using neural rendering techniques such as 3D Gaussian Splatting (3DGS) [45], aiming to build a consistent 3D representation of the scene from RGB observations. These methods often follow a Simultaneous Localization and Mapping (SLAM) paradigm, attempting to jointly reconstruct scene geometry and estimate camera trajectories, similar to pipelines developed for indoor environments. Another direction incorporates monocular depth estimation to provide additional geometric constraints for pose estimation, improving robustness in texture-sparse regions [37]. On the other hand, several studies leverage preoperative imaging data, such as MRI or CT scans, as prior knowledge to guide intraoperative localization [39–41]. These priors can be represented in forms such as tree-structured surgical path graphs

or deformable tissue meshes, enabling alignment between observed endoscopic views and patient-specific anatomical models. Overall, in the low-texture and low-light conditions characteristic of endoscopy, existing methods consistently seek to extract and utilize geometric information as the primary signal for camera localization, either from intraoperative RGB videos or preoperative priors. However, there are still issues with localization accuracy and inference speed that do not meet real-time requirements, as shown in Fig 1(b).

Based on the above observations, we aim to improve camera localization by systematically injecting geometric priors at three complementary scales: structure, motion, and appearance. Such multi-scale modeling has been shown to improve robustness and generalization in related tasks [46]. To this end, we propose a geometry-aware localization framework for bronchoscopic scenarios, which leverages both preoperative and intraoperative geometric information. **Anchor Localization:** We formulate camera localization as an anchor-based coarse-to-fine estimation problem. A graph representation of the airway is constructed from preoperative CT segmentation, where anatomical anchor points are encoded using a Graph Neural Network (GNN). The camera pose is progressively refined from anchor-level matching to precise alignment. **Temporal Motion Tracking:** To improve robustness, we model temporal dynamics using a causal Transformer, enforcing consistency across consecutive frames during training and inference. **Appearance-geometry Matching:** We introduce a cross-modal supervision between RGB observations and rendered depth images from the airway model, enabling geometry-consistent feature learning.

To instantiate this design, we present a Geometry-Aware Bronchoscopy Localization framework (GABL), which enforces explicit geometric supervision across all scales, as illustrated in Fig. 1. The main contributions are summarized as follows.

- We propose a unified geometry-aware bronchoscopy localization framework (GABL) that systematically integrates geometry into localization at three scales, including anchor-based coarse-to-fine localization, temporal motion tracking, and appearance-geometry matching.
- We introduce an anchor-based graph representation derived from preoperative CT as structural geometric priors for bronchoscopy localization, and further incorporate depth-based representations from intraoperative RGB videos to provide complementary appearance-geometry supervision.
- We conduct extensive experiments on a clinically annotated bronchoscopy dataset, demonstrating that our method achieves best localization accuracy while maintaining real-time inference performance.

2 Related Work

2.1 6-DoF Camera Localization

Camera localization aims to estimate the full six-degree-of-freedom (6-DoF) pose of a camera, including its 3D position and orientation. Compared to object localization, this task requires jointly reasoning about translation and rotation, making it inherently more challenging. Existing approaches can be broadly categorized into feature-based methods, learning-based methods, and, more recently, neural rendering-based methods.

Feature-based methods formulate camera localization as a geometric matching problem between 2D image observations and a pre-built 3D scene representation [47]. These approaches typically rely on handcrafted local features to establish 2D–3D correspondences, followed by pose estimation using Perspective-n-Point (PnP) [25, 28, 29, 34, 48]. As a general-purpose solution, such methods achieve strong accuracy in well-textured and stable environments. However, they are sensitive to appearance variations such as illumination changes and motion blur, and often degrade in low-texture or repetitive scenes. In addition, the computational overhead of feature extraction and matching can limit efficiency in large-scale applications.

Learning-based methods estimate camera poses directly from images using deep neural networks, bypassing explicit feature matching and geometric solvers. These approaches can be broadly divided into two categories. Absolute pose regression methods feed images into convolutional or transformer-based networks [20, 22, 44] to predict 6-DoF poses in an end-to-end manner. Scene coordinate regression methods instead predict dense 3D scene coordinates for each pixel [8, 15, 16, 50], which are then used in conjunction with RANSAC and PnP for pose estimation. By leveraging large-scale data, learning-based methods exhibit improved robustness to appearance variations compared to handcrafted features. However, they often suffer from limited generalization to unseen environments and may not achieve the same level of geometric accuracy as feature-based approaches. Recent efforts also explore incorporating additional geometric cues, such as depth or structure information, to further improve robustness.

Neural rendering-based methods have recently emerged alongside advances in 3D scene representation learning [6, 43]. These approaches model the scene as a continuous radiance field and perform camera localization via rendering-based optimization. Specifically, neural implicit representations encode spatial coordinates into color and density, enabling photorealistic novel view synthesis. Many methods follow a SLAM-like paradigm, where Neural Radiance Fields (NeRF) or 3D Gaussian Splatting (3DGS) are used for scene representation, while camera poses are iteratively optimized [17, 26, 30, 33, 49]. Benefiting from strong scene modeling capacity, these methods demonstrate improved robustness in challenging conditions. However, they typically require substantial computational resources, rely on iterative optimization for pose estimation, and are often scene-specific, requiring significant time to adapt to new environments.

In addition, to address challenges in low-texture and low-light environments, graph-based representations have emerged as an effective form of geometric prior. By modeling a scene as a structured graph—where nodes represent keypoints, objects, or semantic regions, and edges encode spatial relationships—localization can be formulated as a graph matching problem, enabling geometric consistency to resolve ambiguities in repetitive structures [35, 42]. Furthermore, pose graph optimization refines camera trajectories by enforcing relative geometric constraints in SLAM systems [19]. Meanwhile, depth information provides direct geometric supervision. RGB-D SLAM methods leverage depth maps for scale-aware motion estimation [23], while recent approaches incorporate depth prediction networks and differentiable geometric losses to enforce

both photometric and geometric consistency [52]. These developments highlight the importance of explicitly modeling geometric structure through either graph connectivity or depth cues for accurate and robust camera localization.

2.2 Endoscope Localization

With the rapid development of surgical robotics and endoscopic imaging systems, camera localization in minimally invasive surgical scenarios has attracted increasing attention [32]. Compared to general-purpose environments, endoscopic localization presents several unique challenges. First, the surgical scene is dominated by deformable soft tissues that undergo continuous non-rigid motion during procedures, making it difficult to maintain consistent geometric correspondence over time. Second, the anatomical workspace is confined and safety-critical, requiring highly accurate and stable pose estimation [11]. Third, endoscopic images often suffer from challenging visual conditions, including limited illumination, specular highlights, and occlusions caused by fluids such as blood or mucus. Furthermore, the lack of distinctive texture in biological tissues significantly degrades the reliability of feature extraction and matching [9].

These factors collectively limit the effectiveness of conventional feature-based and learning-based localization methods, highlighting the need for approaches that better exploit both appearance and geometric information in dynamic and low-texture environments [2].

Most existing datasets are derived from laparoscopic or colonoscopic procedures, such as SCARED [3], StereoMIS [18], EndoMapper [4], EndoSLAM [31], and C3VD [7]. Consequently, localization algorithms for these scenarios are relatively well developed. For example, EndoSLAM [31] introduces Structure-from-Motion (SfM) into endoscopic localization. DARES [37] leverages monocular depth estimation (e.g., Depth Anything v2 [51]) for self-supervised localization. EndoGSLAM [45] incorporates 3D Gaussian Splatting (3DGS) with a streamlined representation to improve efficiency. Endo-FAST3r [36] adopts the foundation model Reloc3r [13] for pose estimation in endoscopic scenes.

In contrast, bronchoscopy localization remains relatively underexplored due to limited data availability. Deng *et al.* [12] construct a dataset from phantom lungs and ex-vivo human lungs, along with a feature-based visual odometry benchmark. PANS [39] and PANSv2 [40] introduce an in-vivo dataset collected from real bronchoscopic procedures and propose a landmark-based localization method. BREATH-VL [41] incorporates visual-language models (VLMs) to enhance localization with semantic information and provides a dataset with preoperative CT and semantic annotations. Moreover, to address data scarcity and annotation challenges, ROOM [14] proposes a framework for generating virtual bronchoscopy datasets from bronchial segmentation, offering a promising direction for data augmentation.

3 Method

In this section, we present the proposed Geometry-Aware Bronchoscopy Localization framework (GABL). As illustrated in Fig. 2, the overall framework consists of three components: geometry prior construction from preoperative airway meshes, multi-modal

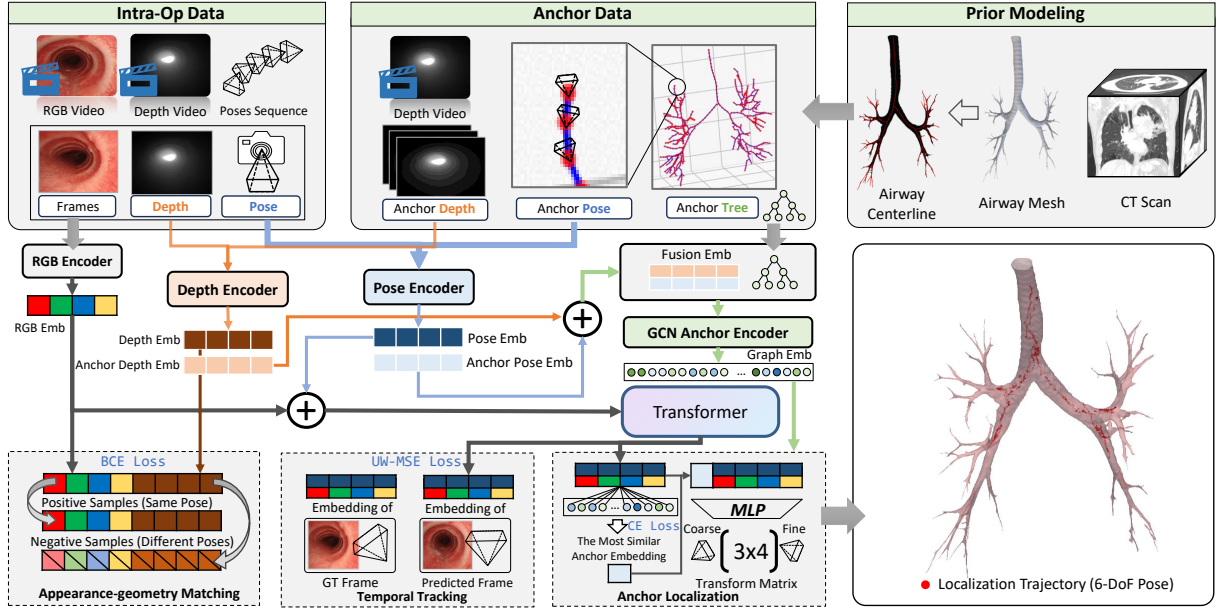


Figure 2: Overview of GABL. The framework takes intraoperative data and anchor-based geometric priors derived from preoperative CT as inputs, and processes them with five representation encoders. It performs multi-stage localization via anchor-based coarse-to-fine estimation, temporal motion tracking, and appearance-geometry alignment, producing a 6-DoF camera trajectory.

representation learning of visual observations and geometric priors, and geometry-aware localization with unified supervision. The camera trajectory is obtained using the in-operative RGB video and the preoperative CT tracheal segmentation mesh.

We first construct a set of anatomical anchor-based geometry priors from the preoperative airway mesh, including a tree-structured representation, predefined 6-DoF poses, and pre-rendered depth maps. We then learn unified representations for both intraoperative observations and preoperative priors using modality-specific encoders, embedding RGB inputs and anchor-related information (e.g., graph structure, pose, and depth) into a shared feature space. Camera localization is performed within a unified framework via an anchor-based coarse-to-fine strategy, progressively refining pose estimates from anchor-level matching to precise alignment. To further improve robustness, we incorporate temporal motion modeling to enforce consistency across frames, and introduce an appearance-geometry alignment mechanism that aligns RGB features with rendered depth representations. All components are jointly optimized, enabling effective integration of structural, temporal, and cross-modal geometric supervision.

3.1 3D Geometric Prior Modeling

To improve training efficiency and reduce computational overhead, we first construct 3D geometric priors offline from preoperative CT images. We use the centerlines extracted from airway models in the dataset as the skeleton representation of the bronchi. Due to the narrow tubular structure of the airways, the centerline closely approximates the bronchoscope trajectory, providing an effective reference for subsequent localization. We uniformly sample a set

of anchor points along the skeleton and construct an anchor graph based on the topological structure of the bronchial tree, ensuring coverage of key anatomical regions such as the carina and multi-level bifurcations.

After determining the anchor point coordinates, we generate corresponding camera poses to simulate the bronchoscope during navigation. Each camera pose is represented as a rigid transformation

$$T = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \in SE(3),$$

where $\mathbf{R} \in SO(3)$ denotes the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ represents the camera center. $SE(3)$ represents the space of rigid body transformations.

We fix the camera center $\mathbf{t}$ at each anchor point and define the target point as its adjacent node in the tree structure. Under this look-at constraint, the viewing direction is uniquely determined, which constrains two degrees of freedom of the rotation $\mathbf{R}$. As a result, the rotation space is reduced from three degrees of freedom to one. The remaining degree corresponds to a rotation around the viewing axis (roll), parameterized by an angle $\theta \in [0, 2\pi)$. We then render depth maps by projecting the 3D geometry under the given camera pose using a rasterization pipeline. Given the camera extrinsics and intrinsics, the depth at each pixel is defined as the distance along the viewing direction to the first visible surface, computed via z-buffering. Compared to Endo-FAST3r [36] and DARES [37], which rely on deep learning models for depth estimation, our approach eliminates the domain gap introduced by learned predictors while maintaining pixel-wise geometric consistency with the camera pose.

In the geometric prior modeling, we have constructed the anchor tree offline, generated camera poses for each anchor, and rendered the corresponding depth maps as geometric priors for supervision, thereby reducing computational cost during training.

3.2 Representation Learning

As illustrated in Fig. 2, we learn a unified embedding space for both intraoperative observations and preoperative geometric priors through a set of modality-specific encoders. The inputs consist of two groups: video observations, including RGB frames, corresponding camera poses, and depth maps; and anchor-related inputs, including the airway anchor graph, predefined anchor poses, and rendered anchor depth maps.

Pose and Depth Encoding. We first encode geometric attributes shared by both inputs. Camera poses and anchor poses are encoded using a multi-layer perceptron (MLP) with three linear layers, producing pose embeddings and anchor pose embeddings, respectively. Similarly, depth maps from both domains are processed by a ResNetV2-18 backbone to obtain depth embeddings and anchor depth embeddings. This design enforces a consistent representation of geometry across intraoperative and preoperative domains.

Visual Encoding. For intraoperative observations, RGB frames are encoded using a ResNetV2-50 backbone to extract visual features, resulting in RGB embeddings. These features capture appearance information complementary to geometric cues.

Anchor Graph Encoding. To capture structural relationships among anchors, we represent the airway anchor tree as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes correspond to anchor points and edges encode their topological connectivity. The graph is treated as undirected to enable bidirectional information propagation.

Each node is initialized with a feature vector by combining anchor pose and depth embeddings:

$$\mathbf{z}_i^{comb(0)} = \mathbf{z}_i^{\text{anchor pose}} + \mathbf{z}_i^{\text{anchor depth}}, \quad (1)$$

We then apply a three-layer graph convolutional network (GCN) to propagate information over the graph. At layer l , node features are updated as:

$$\mathbf{z}_i^{comb(l+1)} = \sum_{j \in \mathcal{N}(i)} \frac{1}{\sqrt{d_i d_j}} \mathbf{W}^{(l)} \mathbf{z}_j^{comb(l)}, \quad (2)$$

where $\mathcal{N}(i)$ denotes the neighborhood of node i , d_i is the node degree, and $\mathbf{W}^{(l)}$ is a learnable transformation.

After three layers, we obtain the final anchor embeddings:

$$\mathbf{z}_i^{\text{anchor}} = \mathbf{z}_i^{comb(3)}, \quad (3)$$

which encode both local geometric attributes and global structural context of the airway tree. This enables each anchor to aggregate information beyond its local neighborhood, capturing the hierarchical structure of the airway.

Temporal Representation Learning. To incorporate motion dynamics, we employ a causal Transformer (pruned from Qwen3 backbone) to model temporal dependencies across video frames. The input to the Transformer is formed by combining RGB embeddings with pose embeddings. Let $\mathbf{z}^{\text{rgb}} \in \mathbb{R}^{n \times s \times d}$ denote the RGB embeddings for a video batch of size n with sequence length s , and $\mathbf{z}^{\text{pose}} \in \mathbb{R}^{n \times s \times d}$ denote the corresponding pose embeddings.

We introduce stochastic masking of pose embeddings as a regularization mechanism. First, a binary mask $\mathbf{m} \in \{0, 1\}^{n \times s}$ is sampled from a Bernoulli distribution:

$$\mathbf{m}_{i,t} \sim \text{Bernoulli}(p), \quad i = 1, \dots, n, \quad t = 1, \dots, s, \quad (4)$$

where $p = 0.25$ is the probability of retaining the original pose embedding. The masked pose embeddings are then defined as:

$$\mathbf{z}_{i,t}^{\text{pose}} = \mathbf{m}_{i,t} \cdot \mathbf{z}_{i,t}^{\text{pose}} + (1 - \mathbf{m}_{i,t}) \cdot \mathbf{z}^{\text{null}}, \quad (5)$$

where $\mathbf{z}^{\text{null}}$ is a learnable null vector representing a missing pose.

The masked pose embeddings are combined with RGB embeddings and fed into the causal Transformer:

$$\mathbf{z}^{\text{video}} = \text{Transformer}(\mathbf{z}^{\text{rgb}} + \mathbf{z}^{\text{pose}}), \quad (6)$$

producing the final video embeddings $\mathbf{z}^{\text{video}}$ that encode both appearance and motion dynamics.

This stochastic pose dropout prevents the model from overfitting to precise pose cues and encourages robust temporal modeling based primarily on RGB observations, effectively regularizing motion dynamics learning.

Overall, this representation learning scheme bridges visual observations and geometric priors in a unified embedding space, providing the foundation for subsequent geometry-aware localization.

3.3 Geometry-aware Localization with Unified Supervision

Anchor Localization. Our model adopts a two-stage anchor-based localization strategy. The coarse localization stage identifies the most relevant anchor as an initial pose estimate, followed by a fine localization stage for refinement.

Given an input frame, we extract its video embedding as $\mathbf{z}_i^{\text{video}}$ from a context window. For a predefined set of anchor points, we denote their embeddings as $\{\mathbf{z}_k^{\text{anchor}}\}_{k=1}^K$.

We compute the similarity logits between the frame embedding and all anchor embeddings:

$$s_{i,k}^{\text{coarse}} = \langle \mathbf{z}_i^{\text{video}}, \mathbf{z}_k^{\text{anchor}} \rangle, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Let y_i denote the ground-truth anchor label for frame i . The coarse localization is formulated as a classification problem using the cross-entropy loss:

$$\mathcal{L}_{\text{coarse}} = - \sum_i \log \frac{\exp(s_{i,y_i}^{\text{coarse}})}{\sum_{k=1}^K \exp(s_{i,k}^{\text{coarse}})}. \quad (8)$$

The predicted anchor index is obtained as:

$$\hat{k}_i = \arg \max_k s_{i,k}^{\text{coarse}}, \quad (9)$$

and the corresponding anchor embedding is taken as the coarse pose representation:

$$\mathbf{z}_i^{\text{coarse}} = \mathbf{z}_{\hat{k}_i}^{\text{anchor}}. \quad (10)$$

For fine-grained localization, the model predicts a refined pose $\hat{\mathbf{T}}_i$ by the Pose Regressor, along with a log-variance term $\log \sigma_{\text{fine},i}^2$ that models the uncertainty of the prediction, following [21].

We adopt an uncertainty-weighted mean squared error (MSE) as the training objective:

$$\mathcal{L}_{\text{fine}} = \frac{1}{N} \sum_i \left(\left\| \hat{\mathbf{T}}_i - \mathbf{T}_i \right\|_2^2 \cdot \exp(-\log \sigma_{\text{fine},i}^2) + \log \sigma_{\text{fine},i}^2 \right), \quad (11)$$

where $\mathbf{T}_i$ denotes the ground-truth pose. This formulation allows the model to adaptively balance regression accuracy and prediction uncertainty.

The proposed two-stage anchor-based localization framework provides several key advantages by effectively leveraging geometric priors. First, the coarse localization stage transforms the continuous pose estimation problem into a discrete anchor classification task, which significantly reduces the search space and stabilizes training. By grounding predictions on predefined anchor points distributed along the airway skeleton, the model benefits from strong global geometric priors. Second, the fine localization stage refines the coarse estimate to achieve precise pose prediction. This coarse-to-fine strategy decouples global localization and local refinement, enabling the model to handle large spatial variations while maintaining high accuracy. Third, the use of structured anchors introduces dense and continuous geometric supervision across the entire trajectory, in contrast to sparse landmark-based methods. This design ensures robust performance even in regions lacking distinctive anatomical landmarks. Finally, the integration of geometric priors at multiple stages improves both translational and rotational accuracy, while effectively mitigating error accumulation over long sequences. Overall, the proposed framework provides a principled and efficient solution for geometry-aware visual localization.

Temporal Motion Tracking. Due to the high frame rate of the bronchoscopic video, the motion between adjacent frames is typically small. Based on this observation, we design a temporal motion tracking module to enforce consistency and reduce drift over time.

Given the video embedding of the current frame $\mathbf{z}_i^{\text{video}}$, the Pose Tracker predicts a relative pose offset $\Delta \mathbf{T}_i$ along with a log-variance term $\log \sigma_{\text{track},i}^2$. The tracked pose is obtained by composing the predicted offset with the previous pose:

$$\hat{\mathbf{T}}_i^{\text{track}} = \mathbf{T}_{i-1} \oplus \Delta \mathbf{T}_i, \quad (12)$$

where $\oplus$ denotes pose composition in $SE(3)$.

We supervise the tracked pose using the ground-truth pose $\mathbf{T}_i$ with an uncertainty-weighted mean squared error loss:

$$\mathcal{L}_{\text{track}} = \frac{1}{N} \sum_i \left(\left\| \hat{\mathbf{T}}_i^{\text{track}} - \mathbf{T}_i \right\|_2^2 \cdot \exp(-\log \sigma_{\text{track},i}^2) + \log \sigma_{\text{track},i}^2 \right). \quad (13)$$

This formulation enforces temporal consistency while allowing the model to adaptively handle motion uncertainty.

Appearance-Geometry Matching. The appearance-geometry matching module aligns RGB and depth features, introducing intra-operative geometric supervision into the localization framework.

Given the RGB embedding of the current frame $\mathbf{z}_i^{\text{rgb}}$ and a sampled depth embedding from intraoperative depth map $\mathbf{z}_j^{\text{depth}}$, we compute the matching logits as $s_{i,j}^{\text{match}}$.

To construct supervision, we adopt a stochastic sampling strategy. With a probability of 50%, we select the corresponding ground-truth depth map ($j = i$) as a positive sample; otherwise, we randomly sample a different frame ($j \neq i$) as a negative candidate.

Instead of using hard binary labels, we define soft labels based on pose similarity:

$$y_{i,j}^{\text{match}} = \exp \left(- \left\| \mathbf{T}_i - \mathbf{T}_j \right\|_2^2 \right), \quad (14)$$

where $\mathbf{T}_i$ and $\mathbf{T}_j$ denote the poses of the current frame and the sampled frame, respectively. This formulation assigns higher similarity scores to geometrically close frames and lower scores to distant ones.

The appearance-geometry matching objective is defined using binary cross-entropy:

$$\mathcal{L}_{\text{match}} = - \sum_i \left[y_{i,j}^{\text{match}} \log \sigma(s_{i,j}^{\text{match}}) + (1 - y_{i,j}^{\text{match}}) \log \left(1 - \sigma(s_{i,j}^{\text{match}}) \right) \right], \quad (15)$$

where $\sigma(\cdot)$ denotes the sigmoid function.

These losses jointly supervise the RGB encoder, pose encoder, depth encoder, anchor graph encoder, and temporal transformer, encouraging consistency between visual appearance and geometric structure.

Inference Strategy. During inference, the model takes only RGB frames as input. The RGB encoder and causal Transformer extract video embeddings, from which the Pose Detector and Pose Tracker independently estimate the camera pose. The detector provides a geometry-aware pose estimate but may exhibit frame-to-frame jitter because of sparse anchors, whereas the tracker produces smoother, temporally consistent estimates but may accumulate errors over time. We therefore select between the two predictions according to their discrepancy. If the discrepancy exceeds an adaptive threshold, the tracked pose is used; otherwise, the detected pose is selected and the threshold is reset. The threshold increases linearly during consecutive tracking steps, encouraging a timely return to detection and preventing long-term drift. This strategy balances geometric accuracy and temporal consistency for robust pose estimation.

4 Experiment

4.1 Datasets and Evaluation Metrics

To evaluate the performance of our model, we conduct experiments on the bronchoscopy localization benchmark BREATH [41]. This dataset contains 66 bronchoscopy procedures, comprising 148,926 frames with 6-DoF pose annotations, along with airway meshes reconstructed from preoperative CT scans.

We adopt Absolute Trajectory Error (ATE) and Success Rate (SR) as evaluation metrics. ATE is decomposed into translational and rotational components: $\text{ATE}_{\text{trans}}$ measures the mean Euclidean distance between predicted and ground-truth camera positions in millimeters, while ATE_{rot} measures the mean angular difference between predicted and ground-truth rotations in degrees. We further report SR-5 and SR-10, defined as the percentages of frames with translational errors below 5 mm and 10 mm, respectively.

Table 1: Performance of 6-DoF localization on Bronchoscope dataset.

Methods	$ATE_{trans}(mm) \downarrow$	$ATE_{rot}(deg) \downarrow$	$SR-5(\%) \uparrow$	$SR-10(\%) \uparrow$
3D Gaussian Splatting				
MonoGS [26]	28.80	105.17	0.34	0.68
EndoGSLAM [45]	26.16	73.59	6.68	19.19
Depth-based Representation				
Endo-FAST3r [36]	22.65	79.15	0.39	1.79
VNB [5]	43.52	109.22	7.06	15.64
DARES [37]	44.58	109.46	2.02	4.32
Landmark Detection				
PANSv2 [40]	10.17	44.49	39.36	67.37
BREATH-VL [41]	7.65	43.32	44.40	75.83
Ours	7.01	29.56	61.04	83.66

4.2 Implementation Details

For 3D geometric prior modeling, we apply farthest point sampling [27] to select 512 anchor points along the airway skeleton, and construct a directed graph based on the topology of the bronchial tree. For each anchor point, two rotation angles in the fixed pose are determined according to the local tangent direction of the skeleton, while the remaining rotation angle is randomly sampled to complete the pose definition. Given the constructed poses, we utilize PyTorch3D to render corresponding depth maps from the airway mesh.

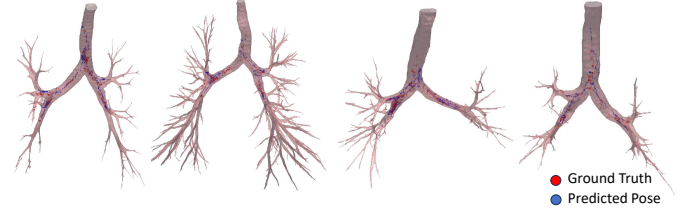
During training, our GABL model is trained on two NVIDIA A100 GPUs, with a per-GPU batch size of 6. Each input video clip contains 64 frames. All input images are resized to 256×256 . The Pose Encoder is implemented as a 3-layer MLP, while the Anchor Encoder is a 3-layer GCN. The RGB Encoder and Depth Encoder adopt ResNetV2-50 and ResNetV2-18 backbones, respectively. The temporal Transformer is adapted from a 24 layers 0.6B Qwen3 model, with a reduced parameter size of 81.82M. All neural network modules in our model use Sigmoid Linear Unit (SiLU) activation functions.

We optimize the model using the AdamW optimizer with a learning rate of 1×10^{-3} , and momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The training losses $\mathcal{L}_{coarse}$, $\mathcal{L}_{fine}$, $\mathcal{L}_{track}$ and $\mathcal{L}_{match}$ have a weighting of 1:1:1:1. The model is trained for 100 epochs.

Across the 66 procedure videos in the dataset, we use 56 cases for training and the remaining 10 cases for testing. The split is performed at the case level to avoid data leakage between training and test sets. To improve data diversity, we employ two data augmentation strategies. When segmenting each input video into clips, we randomly skip 1 to 5 clips to simulate temporal discontinuity. Additionally, each clip has a 50% probability of being temporally reversed, which enhances robustness to motion direction.

4.3 Localization Results

Table 1 presents the localization performance on the BREATH dataset, comparing our method with existing state-of-the-art approaches. The compared methods can be broadly categorized into three groups based on their technical paradigms: 3D Gaussian

**Figure 3: Visualization Results of Our Model**

Splatting (3DGS)-based methods, including MonoGS [26] and EndoGSLAM [45]; depth-based representation methods, such as Endo-FAST3r [36], VNB [5], and DARES [37]; and landmark-based methods, including PANSv2 [40] and BREATH-VL [41].

As shown in Table 1, our GABL achieves a mean ATE_{trans} of 7.01 mm and a mean ATE_{rot} of 29.56° , outperforming the second-best method by 8.37% and 31.76%, respectively. In terms of success rate, our method attains SR-5 of 61.04% and SR-10 of 83.66%, indicating consistently high localization accuracy across frames. The visualization results of our model are shown in Fig. 3.

A closer examination of different method categories highlights the advantages of our geometric design. 3DGS-based methods tend to suffer from error accumulation due to the absence of explicit geometric constraints, leading to degraded localization accuracy over long trajectories. Depth-based methods introduce geometric cues via depth supervision; however, relying on depth alone provides limited structural guidance, which results in suboptimal performance in complex bronchial environments. Landmark-based methods improve localization by incorporating anatomical priors, such as airway carina or semantic landmarks. Nevertheless, these approaches rely on sparse or localized cues, which restrict their effectiveness in regions far from detected landmarks and limit their ability to provide consistent supervision along the full trajectory.

In contrast, our method leverages dense geometric priors by uniformly sampling anchor points along the entire airway skeleton. This design provides global structural coverage, continuous geometric supervision along the trajectory, and robustness in regions lacking distinctive landmarks. By integrating multi-scales

Table 2: Ablation studies of localization components.

Settings	ATE _{trans} ↓	ATE _{rot} ↓	SR-5 ↑	SR-10 ↑
w/o GCN	10.67	31.48	57.59	78.55
w/o Pose Regressor	9.31	107.54	41.09	77.09
w/o Pose Tracker	16.48	38.43	53.57	82.60
w/o Matcher	7.63	30.86	58.09	82.66
Full Model	7.01	29.56	61.04	83.66

Table 3: Ablation study of data augmentation settings.

Settings		Metrics			
Skip	Reverse	ATE _{trans} ↓	ATE _{rot} ↓	SR-5 ↑	SR-10 ↑
-	-	7.97	31.71	54.75	80.72
✓	-	7.84	32.24	58.02	82.08
-	✓	7.71	33.56	59.65	81.52
✓	✓	7.01	29.56	61.04	83.66

Table 4: Ablation study of dropout rate in causal Transformer.

Dropout rate	ATE _{trans} ↓	ATE _{rot} ↓	SR-5 ↑	SR-10 ↑
0.25	8.04	31.51	57.94	83.12
0.50	7.59	30.76	58.27	82.74
0.75	7.01	29.56	61.04	83.66
1.0	7.94	31.25	58.57	82.48

of geometric constraints, our approach effectively reduces drift and improves both translational and rotational accuracy, leading to superior overall localization performance.

We evaluate the inference speed on a consumer-grade NVIDIA RTX 4090 GPU. Our model achieves 33.6 FPS, significantly outperforming PANSv2 [40] and BREATH-VL [41], which run at 8.5 FPS and 5.6 FPS, respectively. In comparison, 3DGS-based methods such as EndoGSLAM [45] exhibit lower efficiency in this scenario, achieving only 2.7 FPS in our experiments. This is likely due to the increased complexity of the airway lumen, which poses challenges for stable mapping and rendering. Overall, our method achieves a 4× to 12× speedup over prior approaches, while maintaining superior localization accuracy. This level of efficiency satisfies the real-time requirements of clinical endoscopy.

4.4 Ablation Studies

To validate the effectiveness of our geometry-aware design, we conduct ablation studies on four key components: the Anchor Encoder, Pose Regressor for fine-grained localization, Pose Tracker for temporal modeling, and Matcher for appearance-geometry alignment.

We first evaluate the contribution of each module by removing or modifying it individually. Specifically, we replace the GCN in the Anchor Encoder with a linear layer and remove the anchor tree input, thereby eliminating structured geometric priors from preoperative data. Removing the Pose Regressor means directly using coarse localization results as the final output. Disabling the

Pose Tracker removes temporal motion supervision. Removing the Matcher prevents the model from learning the correspondence between RGB images and rendered depth maps, thus discarding geometry-aware representation learning.

As shown in Table 2, removing any component leads to performance degradation, demonstrating that each module contributes to the overall system. In particular, removing the Pose Tracker increases the mean ATE_{trans} by 9.47 mm, indicating that temporal geometric constraints are essential to prevent error accumulation along the trajectory.

Moreover, removing the Pose Regressor significantly degrades rotational accuracy, with ATE_{rot} increasing to 107.54°. This suggests that coarse localization alone provides limited orientation information, and fine-grained geometric refinement is necessary for accurate pose estimation.

When the anchor tree structure is removed and the GCN is replaced, the mean ATE_{trans} increases to 10.67 mm. This result demonstrates that, beyond serving as coarse localization references, the tree-structured airway representation provides meaningful global geometric priors that improve localization accuracy.

As discussed previously, we employ two data augmentation strategies: random clip skipping and temporal reversal. We further analyze their impact in Table 3. These augmentations reduce the mean ATE_{trans} by 0.70 to 0.96 mm. Random clip skipping exposes the model to temporally discontinuous inputs, improving robustness to irregular motion. Temporal reversal simulates backward navigation of the bronchoscope, enabling the model to better handle bidirectional motion patterns commonly observed in practice.

We further investigate the effect of pose dropout in the temporal Transformer. Specifically, we apply a Bernoulli distribution to randomly mask pose inputs for each frame, which prevents the model from over-relying on explicit pose signals. As shown in Table 4, a dropout rate of 0.75 achieves the best performance, indicating improved robustness to motion variations. In contrast, lower dropout rates result in a performance drop of 0.58 to 1.03 mm in ATE_{trans}, suggesting insufficient regularization. When the dropout rate is set to 1 (i.e., pose information is completely discarded), the model’s performance declines, suggesting that combining pose information with RGB data is more effective for motion supervision.

5 Conclusion

In this work, we present GABL, a geometry-aware localization framework for bronchoscopy that systematically integrates geometric priors at three complementary scales: structure, motion, and appearance. By combining anchor-based coarse-to-fine localization, temporal motion modeling, and cross-modal appearance-geometry supervision, our method unifies preoperative and intraoperative information within a single framework. Extensive experiments on the BREATH dataset demonstrate that GABL achieves a mean ATE_{trans} of 7.01 mm and a mean ATE_{rot} of 29.56°, with a success rate (SR-5) of 61.04%, outperforming existing methods while maintaining real-time inference performance. These results highlight the effectiveness of multi-scale geometric constraints for robust camera localization in challenging endoscopic environments and suggest promising potential for broader applications in medical and robotic navigation. Code is available online.

Acknowledgments

This work was supported by the InnoHK initiative of the Government of the Hong Kong Special Administrative Region.

References

- [1] June Hong Ahn. 2020. An update on the role of bronchoscopy in the diagnosis of pulmonary disease. *Yeungnam University Journal of Medicine* 37, 4 (2020), 253–261.
- [2] Sharib Ali. 2022. Where do we stand in AI for endoscopic image analysis? Deciphering gaps and future directions. *npj Digital Medicine* 5, 1 (2022), 184.
- [3] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al. 2021. Stereo correspondence and reconstruction of endoscopic data challenge. *arXiv preprint arXiv:2101.01133* (2021).
- [4] Pablo Azagra, Carlos Sostres, Ángel Ferrández, Luis Riazuelo, Clara Tomasini, O León Barbed, Javier Morlana, David Recasens, Victor M Batlle, Juan J Gómez-Rodríguez, et al. 2023. Endomapper dataset of complete calibrated endoscopy procedures. *Scientific Data* 10, 1 (2023), 671.
- [5] Artur Banach, Franklin King, Fumitaro Masaki, Hisashi Tsukada, and Nobuhiko Hata. 2021. Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation. *Medical image analysis* 73 (2021), 102164.
- [6] Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo. 2025. 3D Gaussian Splatting: Survey, Technologies, Challenges, and Opportunities. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 7 (2025), 6832–6852. doi:10.1109/TCSVT.2025.3538684
- [7] Taylor L Bobrow, Mayank Golhar, Rohan Vijayan, Venkata S Akshintala, Juan R Garcia, and Nicholas J Durr. 2023. Colonoscopy 3D video dataset with paired depth from 2D-3D registration. *Medical image analysis* 90 (2023), 102956.
- [8] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. 2017. DSAC - Differentiable RANSAC for Camera Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [9] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian Jun Zhang. 2018. SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality. *Computer methods and programs in biomedicine* 158 (2018), 135–146.
- [10] Joseph Cicienia, Sameer K Avasarala, and Thomas R Gildea. 2020. Navigational bronchoscopy: a guide through history, current use, and developing technology. *Journal of thoracic disease* 12, 6 (2020), 3263.
- [11] Kristoffer Mazanti Cold, Sujun Xie, Anne Orholm Nielsen, Paul Frost Clementsen, and Lars Konge. 2024. Artificial intelligence improves novices' bronchoscopy performance: a randomized controlled trial in a simulated setting. *Chest* 165, 2 (2024), 405–413.
- [12] Janning Deng, Peize Li, Kevin Dhaliwal, Chris Xiaoxuan Lu, and Mohsen Khadem. 2023. Feature-based Visual Odometry for Bronchoscopy: A Dataset and Benchmark. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 6557–6564. doi:10.1109/IROS55552.2023.10342034
- [13] Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. 2025. Reloc3r: Large-Scale Training of Relative Camera Pose Regression for Generalizable, Fast, and Accurate Visual Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16739–16752.
- [14] Salvatore Esposito, Matias Mattamala, Daniel Rebain, Francis Xiatian Zhang, Kevin Dhaliwal, Mohsen Khadem, and Subramanian Ramamoorthy. 2025. ROOM: A Physics-Based Continuum Robot Simulator for Photorealistic Medical Datasets Generation. *arXiv preprint arXiv:2509.13177* (2025).
- [15] Khang Truong Giang, Soohwan Song, and Sungho Jo. 2024. Learning to Produce Semi-dense Correspondences for Visual Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19468–19478.
- [16] Abner Guzman-Rivera, Pushmeet Kohli, Ben Glocker, Jamie Shotton, Toby Sharp, Andrew Fitzgibbon, and Shahram Izadi. 2014. Multi-Output Learning for Camera Relocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [17] Seongho Ha, Jiung Yeon, and Hyeonwoo Yu. 2024. Rgbd gs-icp slam. In *European conference on computer vision*. Springer, 180–197.
- [18] Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. 2023. Learning how to robustly estimate camera pose in endoscopic videos. *International journal of computer assisted radiology and surgery* 18, 7 (2023), 1185–1192.
- [19] Andela Jurić, Filip Kendeš, Ivan Marković, and Ivan Petrović. 2021. A comparison of graph optimization approaches for pose estimation in SLAM. In *2021 44th international convention on information, communication and electronic technology (MIPRO)*. IEEE, 1113–1118.
- [20] Alex Kendall and Roberto Cipolla. 2017. Geometric Loss Functions for Camera Pose Regression With Deep Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [21] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [22] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [23] Christian Kerl, Jürgen Sturm, and Daniel Cremers. 2013. Dense visual SLAM for RGB-D cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2100–2106. doi:10.1109/IROS.2013.6696650
- [24] Hu Lin, Chengjiang Long, Yifeng Fei, Qianchen Xia, Erwei Yin, Baoai Yin, and Xin Yang. 2024. Exploring matching rates: From keypoint selection to camera relocalization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 506–514.
- [25] Liu Liu, Hongdong Li, and Yuchao Dai. 2017. Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [26] Hidenobu Matsuki, Riku Murai, Paul HJ Kelly, and Andrew J Davison. 2024. Gaussian splatting slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18039–18048.
- [27] Carsten Moenning and Neil A. Dodgson. 2003. *Fast Marching farthest point sampling*. Technical Report UCAM-CL-TR-562. University of Cambridge, Computer Laboratory. doi:10.48456/tr-562
- [28] L. Morelli, F. Ioli, R. Beber, F. Menna, F. Remondino, and A. Vitti. 2023. COLMAP-SLAM: A FRAMEWORK FOR VISUAL ODOMETRY. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLVIII-1/W1-2023* (2023), 317–324. doi:10.5194/isprs-archives-XLVIII-1-W1-2023-317-2023
- [29] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. 2016. Open-MVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*. Springer, 60–74.
- [30] Riku Murai, Eric Dexeimer, and Andrew J. Davison. 2025. MAST3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16695–16705.
- [31] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfiz Coskun, Kagan Incecan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al. 2021. EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. *Medical image analysis* 71 (2021), 102058.
- [32] Laura Privitera, Irene Paraboschi, Divyansh Dixit, Owen J Arthurs, and Stefano Giuliani. 2022. Image-guided surgery and novel intraoperative devices for enhanced visualisation in general and paediatric surgery: A review. *Innovative Surgical Sciences* 6, 4 (2022), 161–172.
- [33] Antoni Rosinol, John J Leonard, and Luca Carlone. 2023. Nerf-slam: Real-time dense monocular slam with neural radiance fields. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3437–3444.
- [34] Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, and Torsten Sattler. 2021. Back to the Feature: Learning Robust Camera Localization From Pixels To Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3247–3257.
- [35] Torsten Sattler, Akihiko Torii, Josef Sivic, Marc Pollefeys, Hajime Taira, Masatoshi Okutomi, and Tomas Pajdla. 2017. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [36] Mona Sheikh Zeinoddin, Mobarak I Hoque, Zafer Tandogdu, Greg L Shaw, Matthew J Clarkson, Evangelos B Mazomenos, and Danail Stoyanov. 2025. Endo-fast3r: endoscopic foundation model adaptation for structure from motion. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 117–126.
- [37] Mona Sheikh Zeinoddin, Chiara Lena, Jiongqi Qu, Luca Carlini, Mattia Magro, Seunghoi Kim, Elena De Momi, Sophia Bano, Matthew Grech-Sollars, Evangelos Mazomenos, et al. 2024. Dares: Depth anything in robotic endoscopic surgery with self-supervised vector-lora of the foundation model. In *European Conference on Computer Vision*. Springer, 1–11.
- [38] Trung Duc Than, Gursel Alici, Hao Zhou, and Weihua Li. 2012. A Review of Localization Systems for Robotic Endoscopic Capsules. *IEEE Transactions on Biomedical Engineering* 59, 9 (2012), 2387–2399. doi:10.1109/TBME.2012.2201715
- [39] Qingyao Tian, Zhen Chen, Huai Liao, Xinyan Huang, Bingyu Yang, Lujie Li, and Hongbin Liu. 2024. PANS: probabilistic airway navigation system for real-time robust bronchoscope localization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 466–476.
- [40] Qingyao Tian, Huai Liao, Xinyan Huang, Bingyu Yang, and Hongbin Liu. 2025. Harnessing Foundation Models for Robust and Generalizable 6-DOF Bronchoscopy Localization. In *International Workshop on Agentic AI for Medicine*. Springer, 127–135.

- [41] Qingyao Tian, Bingyu Yang, Huai Liao, Xinyan Huang, Junyong Li, Dong Yi, and Hongbin Liu. 2026. BREATHE-VL: Vision-Language-Guided 6-DoF Bronchoscopy Localization via Semantic-Geometric Fusion. *arXiv:2601.03713 [cs.CV]* <https://arxiv.org/abs/2601.03713>
- [42] Mehmet Ozgur Turkoglu, Eric Brachmann, Konrad Schindler, Gabriel J. Brostow, and Aron Monszpart. 2021. Visual Camera Re-Localization Using Graph Neural Networks and Relative Pose Supervision. In *2021 International Conference on 3D Vision (3DV)*. 145–155. doi:10.1109/3DV53792.2021.00025
- [43] Guangming Wang, Lei Pan, Songyou Peng, Shaohui Liu, Chenfeng Xu, Yanzi Miao, Wei Zhan, Masayoshi Tomizuka, Marc Pollefeys, and Hesheng Wang. 2024. NeRFs in robotics: A survey. *The International Journal of Robotics Research* (2024), 02783649251374246.
- [44] Jianyuan Wang, Christian Rupprecht, and David Novotny. 2023. PoseDiffusion: Solving Pose Estimation via Diffusion-aided Bundle Adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9773–9783.
- [45] Kailing Wang, Chen Yang, Yuehao Wang, Sikuang Li, Yan Wang, Qi Dou, Xiaokang Yang, and Wei Shen. 2024. Endogslam: Real-time dense reconstruction and tracking in endoscopic surgeries using gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 219–229.
- [46] Yuntao Wang, Jinpu Zhang, Ruonan Wei, Wenbo Gao, and Yuehuan Wang. 2024. Mfrgn: Multi-scale feature representation generalization network for ground-to-aerial geo-localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2574–2583.
- [47] Yihong Wu, Fulin Tang, and Heping Li. 2018. Image-based camera localization: an overview. *Visual Computing for Industry, Biomedicine, and Art* 1, 1 (2018), 8.
- [48] Tao Xie, Kun Dai, Ke Wang, Ruifeng Li, Jiahe Wang, Xinyue Tang, and Lijun Zhao. 2022. A Deep Feature Aggregation Network for Accurate Indoor Camera Localization. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3687–3694. doi:10.1109/LRA.2022.3146946
- [49] Chi Yan, Delin Qu, Dan Xu, Bin Zhao, Zhigang Wang, Dong Wang, and Xuelong Li. 2024. GS-SLAM: Dense Visual SLAM with 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 19595–19604.
- [50] Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, and Ping Tan. 2019. SANet: Scene Agnostic Network for Camera Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [51] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything V2. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 21875–21911. doi:10.52202/079017-0688
- [52] Huangying Zhan, Chamara Saroj Weerasekera, Jia-Wang Bian, Ravi Garg, and Ian Reid. 2021. DF-VO: What should be learnt for visual odometry? *arXiv preprint arXiv:2103.00933* (2021).