Towards Robust Bronchoscopic SLAM: Transformer-Based Monocular Depth for Gaussian Splatting

Justin Yee[1], Haoran Zhou [2], Raj Priyadarshi[2], Samyu Kamtam[2], Taeyoung Kang [3], Hongyi Li [3], Rui Li[1*]

***Abstract*— Bronchoscopic Simultaneous Localization and Mapping (SLAM) is a critical technology for autonomous navigation in minimally invasive lung interventions. Traditional SLAM methods face challenges in highly homogeneous airway environments due to textureless surfaces, lack of depth estimation, and limited visual features. In this work, we propose a transformer-based monocular depth estimation approach tailored for bronchoscopic navigation and integrate it into a Gaussian splatting based SLAM. Our method additionally leverages deep neural networks to infer depth from bronchoscopic photometry and depth estimation using photo edge blurriness, enabling robust 3D reconstruction while maintaining computational efficiency. By incorporating Gaussian Splatting, we achieve high-fidelity mapping with adaptive uncertainty modeling, improving localization accuracy and scene representation. Experimental results demonstrate our preliminary result in challenging airway environments, advancing more reliable and autonomous bronchoscopic navigation.**

I. Introduction

Accurate localization during bronchoscopy is essential for navigating complex airway structures. Existing monocular bronchoscopy methods often rely on preoperative CT scans or external tracking [1], which limits adaptability and real-time applicability. Purely visual-based navigation offers a potential solution, with depth estimation playing a crucial role in enabling SLAM for the lung environment. However, existing depth estimation models such as Depth Anything [2] performs poorly in this setting due to challenges such as low-texture regions and the complex geometry of airway tunnels. To address these limitations, we propose a monocular bronchoscopy model that aggregates photometric cues and edge blurriness to enhance depth estimation. This refined depth information is then integrated into a modified Gaussian Splatting SLAM framework [3] designed for bronchoscopy, improving both depth accuracy and localization performance for more reliable bronchoscopic navigation.

II. Method

Depth Anything [2] is a Vision Transformer-based depth estimation model that builds upon the DINOv2 [4] and DPT architectures [5], using the DINOv2 encoder for feature extraction and the DPT decoder to refine depth predictions. This work extends the Depth Anything V2-Base foundational model, fine tuning it on a labeled stereoscopic bronchoscopy dataset captured using two ES101 monocular endoscopes from Shenzhen VisionMeta Technology Co. Additionally, we incorporate an image blurriness-based depth estimation method and the photometric-based depth estimation technique from MonoLoT [6] as physical constraints.

The method of estimating depth using blurriness is based on the fixed focal length of the endoscope. If the object's depth from the camera equals the focal length, it is at the focus and is least blurry. Otherwise, if the distance to the camera is smaller than the focal length, the object's blurriness increases. To assess blurriness numerically, our method evaluates the output of a Fast Fourier Transform. Thus, the numerical relationship between blurriness and depth can be constructed after calibration.

$$\hat{d} = D(\bar{b}) \tag{1}$$

In equation 1, $\hat{d}$ is the estimated depth, $\bar{b}$ is the calculated blurriness, and $D(\bar{b})$ is the calibrated depth function. The parameters θ^* of $D(\bar{b})$ are derived from n data points of (b_i, d_i) collected in the calibration and follow the approach below in equation 2.

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{n} \left(d_i - D(b_i; \theta)\right)^2 \tag{2}$$

MonoLoT [6] proposes an extension to the state-of-the-art framework from DepthNet [7] and PoseNet [8], which calculates the inverse depth and viewpoint of the image pair (I_a, I_b). MonoLoT optimizes the original loss function (3), by adding two additional parameters, point matching loss $\mathcal{L}_m$ and batch shuffle loss $\mathcal{L}_{bis.}$, shown in equation 4. In both equations $\mathcal{L}_r$ is the minimum reprojection loss and $\mathcal{L}_s$ is the edge-aware smoothness loss, with μ used for auto-masking and λ a hyperparameter

$$\mathcal{L} = \mu\mathcal{L}_r + \lambda\mathcal{L}_s. \tag{3}$$

$$\mathcal{L} = \mu\mathcal{L}_r + \lambda\mathcal{L}_s + \mathcal{L}_m + \mathcal{L}_{bis}. \tag{4}$$

Image data captured from a bronchoscopy phantom (Figure 1) by a ES101 was fed into the three algorithms to perform real-time depth estimation, the results aggregated to improve precision.

To evaluate the effectiveness of these additional physical constraints and stereoscopic data, we performed an ablation study to assess their impact on the depth estimation performance.

The resulting data from the monocular depth estimation is then used by our Gaussian Splatting SLAM to perform real-time bronchoscopic localization and scene reconstruction. For visualization, 3D Gaussian Splatting (3DGS) [3] is used to map the scene using a set of anisotropic Gaussians G_i, where each Gaussian is characterized by its color c_i and

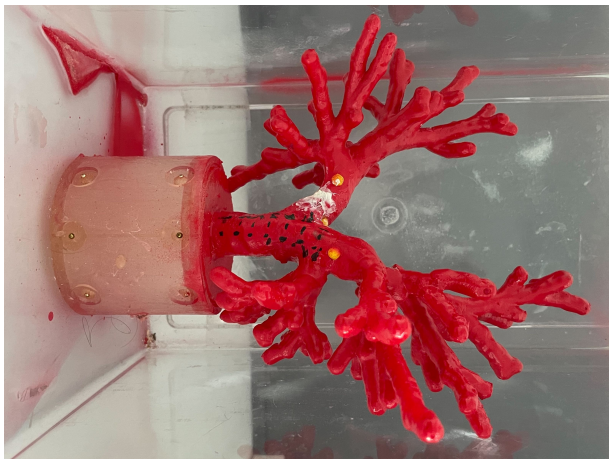

Fig. 1: Bronchoscopy phantom

opacity α_i. The mean $\mu_i W$ and covariance $\sum iW$ of each Gaussian, defined in world coordinates, represent the position and shape of the splat in 3D space. The mean indicates the location of each splat, and the covariance defines its shape, allowing it to be spherical or elliptical. For simplicity we do not incorporate spherical harmonics for view-dependent radiance in this work.

3DGS is especially useful in non-constrained spaces, such as inside the human body, where the geometry of the scene can vary greatly. Since endoscopic images often deal with complex, flexible, and unstructured environments, volume rendering is ideal as it doesn't require explicit surface extraction. Instead, by splatting and blending N Gaussians, we can synthesize the pixel color C_p as follows:

$$C_p = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{5}$$

Equation 5 ensures that overlapping Gaussians blend smoothly, with each splat contributing its color and opacity to the final pixel. One of the key advantages of this method is that it can be used both with and without depth estimation. However, when depth estimation is available, it provides additional information about the scene geometry, which can be used to refine the positioning and rendering of the Gaussians. Depth information can improve the accuracy of the splatting process, leading to a more precise 3D reconstruction of the scene.

To project the 3D Gaussians in world coordinates onto the 2D image plane, we use a projective transformation:

$$\mu_i = \pi(T_{CW} \cdot \mu_w), \Sigma_I = JW\Sigma_W W^T J^T \tag{6}$$

where π is the projection operation and $T_C W$ is the camera pose of the viewpoint. J is the Jacobian of the linear approximation of the projective transformation and W is the rotational component of camera pose.

III. Results and Discussions

We evaluated our depth estimation model and Gaussian Splatting SLAM using a bronchoscopy phantom. Fig. 4 shows the current reconstructed airway structure. By incorporating photometric depth estimation and edge blurriness analysis, our method generates a consistent depth map from monocular endoscopic images. The Gaussian Splatting SLAM framework effectively utilizes this depth map to create a smooth and continuous airway map.

Fig. 4: Preliminary reconstruction results

Our model performs well in structured airway regions but faces challenges with severe illumination changes and hand tremors. Small depth estimation errors can accumulate over SLAM iterations, causing slight localization drift. Furthermore, our current bronchoscopy model is rigid and unable mimic the respiratory behavior of real bronchi. Future work will focus on improving depth estimation, improving SLAM stability, and validating the framework in real dynamic bronchoscopic procedures.

References

[1] E. T. Sumner, J. Chang, P. R. Patel, H. Bedi, and B. D. Shaller, "State of the art: peripheral diagnostic bronchoscopy," *Journal of Thoracic Disease*, vol. 16, no. 8, p. 5409, 2024.

[2] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 371–10 381.

[3] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 18 039–18 048.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[5] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.

[6] Q. He, G. Feng, S. Bano, D. Stoyanov, and S. Zuo, "Monolot: Self-supervised monocular depth estimation in low-texture scenes for automatic robotic endoscopy," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 10, pp. 6078–6091, 2024.

[7] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Depthnet: A recurrent neural network architecture for monocular depth prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 283–291.

[8] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.

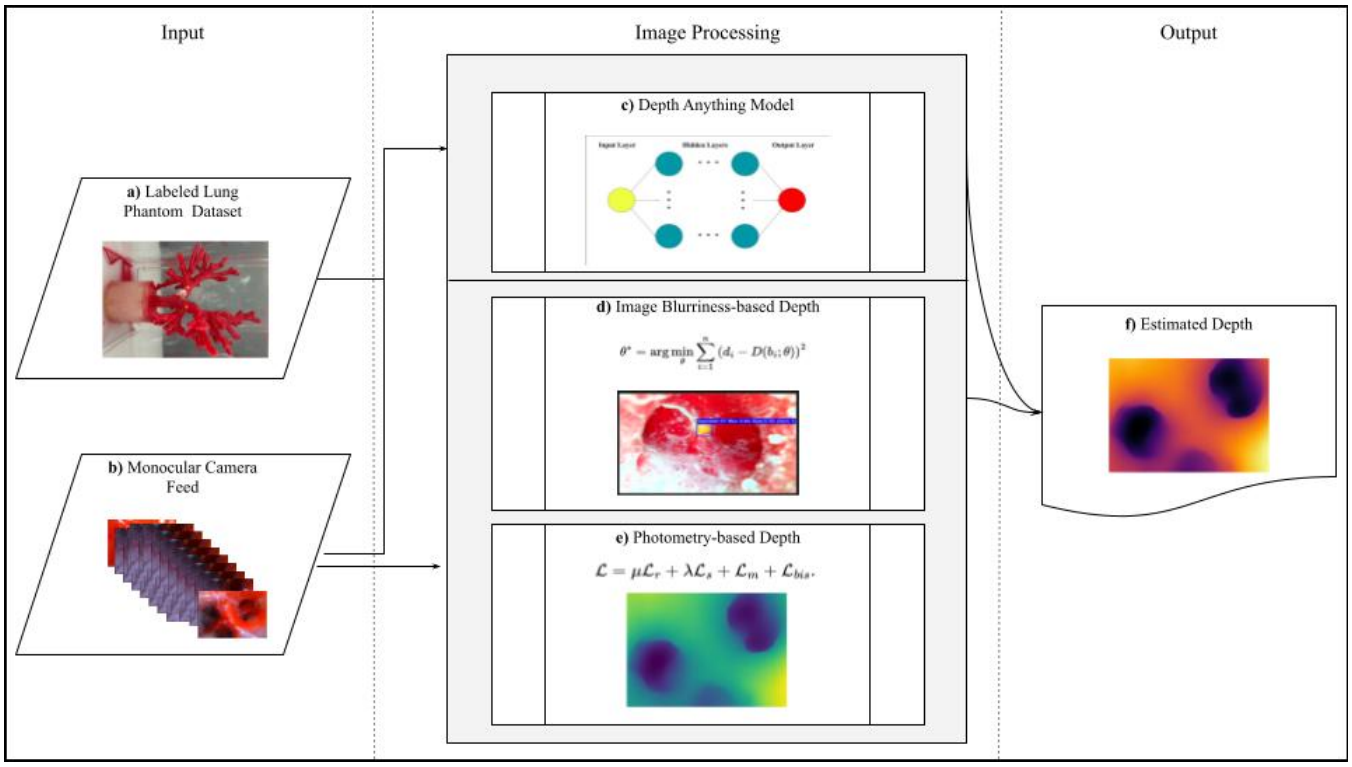

Fig. 2: Proposed robust depth estimation. The training dataset for Depth Anything is collected from the lung phantom shown in a). The monocular endoscope feed is illustrated in b). A general network diagram for Depth Anything is shown in c). d) displays the depth obtained from image blurriness, e) displays the depth obtained from camera photometry, f) displays the final aggregated depth map.

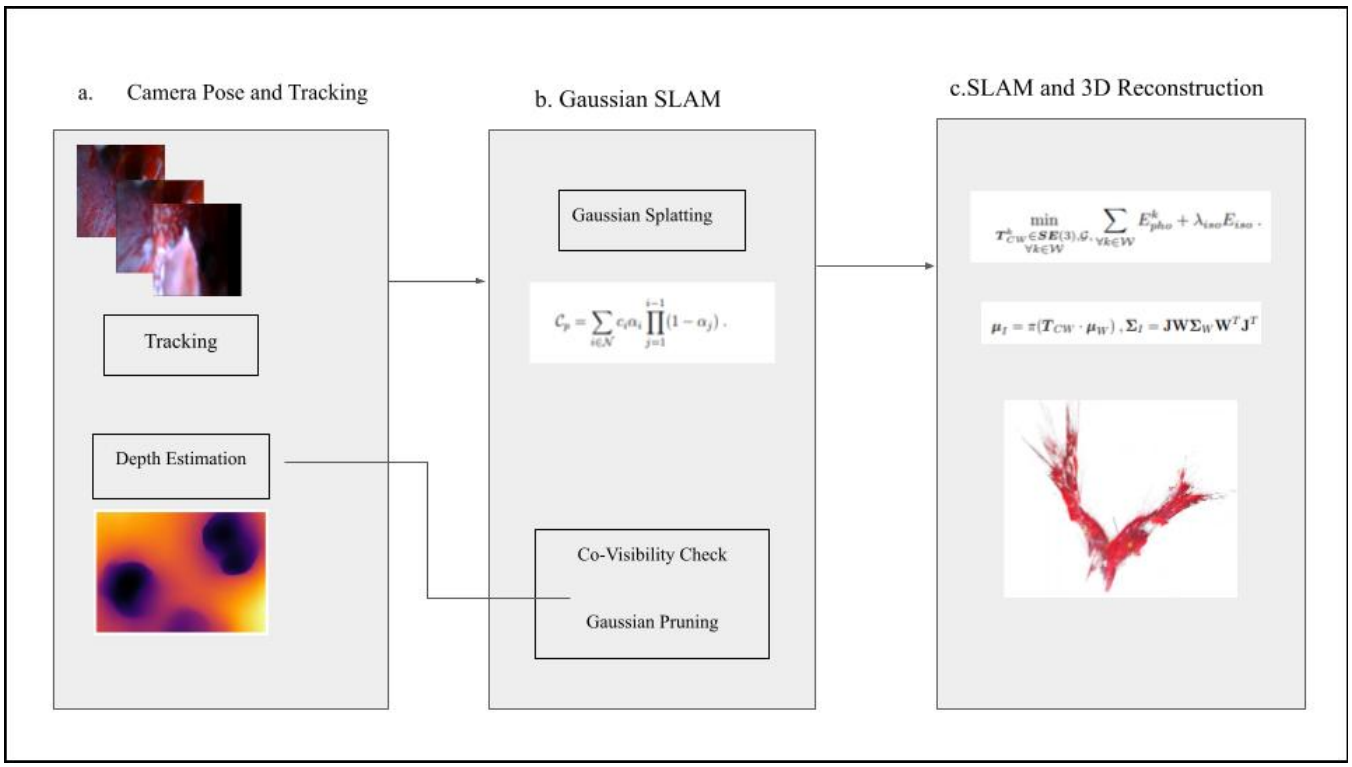

Fig. 3: Proposed Gaussian splatting SLAM workflow. a) depicts the camera pose and tracking branch, with monocular endoscope image input and depth input acquired from the previous depth pipeline, b) depicts the Gaussian Splatting branch, c) depicts the 3D reconstruction branch and preliminary 3D reconstruction