2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871588

Endoscope Localization and Dense Surgical Scene Reconstruction for Stereo Endoscopy by Unsupervised Optical Flow and Kanade-Lucas-Tomasi Tracking

Zixin Yang[1], Shan Lin[2], Richard Simon[3], and Cristian A. Linte[1,3]

Abstract— **In image-guided surgery, endoscope tracking and surgical scene reconstruction are critical, yet equally challenging tasks. We present a hybrid visual odometry and reconstruction framework for stereo endoscopy that leverages unsupervised learning-based and traditional optical flow methods to enable concurrent endoscope tracking and dense scene reconstruction. More specifically, to reconstruct texture-less tissue surfaces, we use an unsupervised learning-based optical flow method to estimate dense depth maps from stereo images. Robust 3D landmarks are selected from the dense depth maps and tracked via the Kanade-Lucas-Tomasi tracking algorithm. The hybrid visual odometry also benefits from traditional visual odometry modules, such as keyframe insertion and local bundle adjustment. We evaluate the proposed framework on endoscopic video sequences openly available via the SCARED dataset against both ground truth data, as well as two other state-of-the-art methods - ORB-SLAM2 and Endo-depth. Our proposed method achieved comparable results in terms of both RMS Absolute Trajectory Error and Cloud-to-Mesh RMS Error, suggesting its potential to enable accurate endoscope tracking and scene reconstruction.**

Index Terms— **Stereo Endoscopy, Visual Odometry, Surgical Scene Reconstruction.**

I. INTRODUCTION

Endoscope tracking is an essential component of image-guided interventions that rely on video views for surgical instrument navigation. Both optical (OTS) and electromagnetic tracking systems (EMT) could be used to track an endoscope [1]. However, despite their high tracking accuracy, OTS require direct line-of-sight between the tracker and a dynamic reference frame rigidly mounted on the endoscope; similarly, EMTs are susceptible to magnetic field distortion from surrounding metal or other ferromagnetic sources. An alternative approach is to track the endoscope using only the image data captured by the endoscope, a method also known as visual odometry (VO) [2]. This approach is appealing, as it mitigates the limitations associated with external tracking, while requiring minimum modifications to the existing surgical workflow.

Endoscope tracking is also paramount for surgical scene reconstruction [3], [4], which entails the fusion of dense depth maps and color images with estimated camera poses [5]. The reconstruction benefits many downstream tasks, including visual analysis [4] and image-to-patient registration [1]. Hence, given the critical need for accurate endoscope tracking to achieve faithful surgical scene reconstruction, in this work, we target both visual odometry and scene reconstruction.

Existing methods for camera localization and dense scene reconstruction include traditional methods based on multi-view constraints [6], [7], end-to-end deep learning methods [8], and hybrid methods [3], [4] that replace several modules of traditional methods with deep learning methods. Traditional approaches using multi-view correspondences and constraints can yield highly accurate results of tracking and reconstruction given well-textured images. However, correspondences are difficult to estimate from texture-less surfaces, which are typical in endoscopic images, resulting in sparse reconstruction.

Deep learning-based dense scene reconstruction methods have shown promising results, especially for dense depth estimation. Ozyoruk *et al.* [8] used an end-to-end deep learning method based on a depth estimation network and a pose estimation network. However, end-to-end deep learning methods lack bundle adjustment[7], [2], which leads to the accumulation of tracking drift [4]. Moreover, when the training data and testing data have different data distributions, end-to-end deep learning methods may also suffer from domain gaps [9]. Hence, hybrid methods [3], [4] that leverage the power of both traditional and deep learning methods have shown further potential. One such example is the work by Recasens *et al.* [3], which employs self-supervised depth networks to generate pseudo-RGBD frames, then track the camera using photometric constraints.

To further mitigate the limitations associated with accumulated tracking drift and sparse reconstruction, here we propose a hybrid visual odometry and dense scene reconstruction framework (END-VO). New techniques in END-VO include: 1) An unsupervised learning-based optical flow method [10] is employed to estimate dense depth maps from low-textured stereo endoscopic images. 2) To accurately estimate camera poses, we design a rule to select accurate and easy-to-track landmarks from both Kanade-Lucas-Tomasi (KLT) tracking [11] and unsupervised optical flow [10]. 3) We leverage traditional modules to improve the tracking and the reconstruction performance, such as bundle adjustment to reduce the accumulation of tracking drift, and keyframe insertion module to prevent significant point cloud overlap and ensure accurate tracking. We evaluate

[1]Center for Imaging Science, Rochester Institute of Technology Rochester, NY 14623, USA yy8898@g.rit.edu
[2]Electrical and Computer Engineering, University of California San Diego, CA 92093, USA
[3]Biomedical Engineering, Rochester Institute of Technology, USA, NY 14623, USA calbme@rit.edu

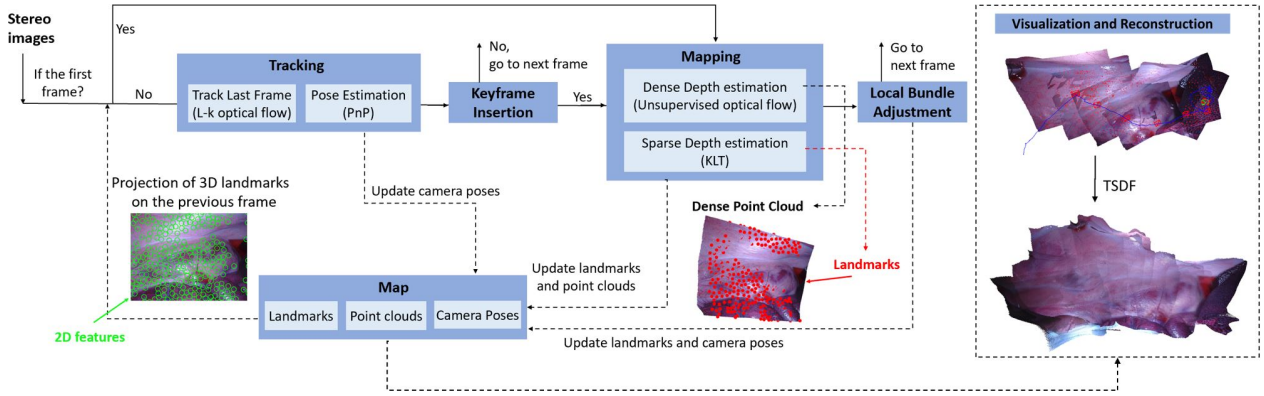


Fig. 1. END-VO overview, showing main modules: 1) Mapping - predicts dense depth map and selects robust 3D landmarks for tracking; 2) Tracking - infers current camera pose relative to previous pose; 3) Key Frame Insertion - triggers mapping and performs local bundle adjustment, if needed; 4) Local Bundle Adjustment - optimizes existing landmarks and key frame camera poses; 5) Visualization and Reconstruction - displays the depth map and builds a global 3D mesh model. Detected landmarks, reconstructed point clouds, and estimated camera poses are all stored in the Map. Green and red represent 2D features and 3D landmarks, respectively.

our proposed framework on the Stereo Correspondence and Reconstruction of Endoscopic (SCARED) dataset [12] and showed competitive results, comparable to those yielded by other state-of-the-art methods.

II. Methods

A. END-VO Framework

Our method uses a sequence of stereo endoscopic videos with known camera calibration parameters as input to track the stereo endoscope, and reconstructs the surface as a global 3D mesh model. The proposed framework comprises several modules illustrated in Fig. 1 and also described below. The framework gradually builds a map based on 3D landmarks, point clouds, and camera poses. Camera poses are estimated using 3D landmarks and their associated 2D image features. Our algorithm updates the 3D landmarks and point clouds of the map only when a new keyframes is inserted [2], [7]. Finally, a global mesh model is constructed using the dense point clouds and associated camera poses.

1) Mapping: This module updates the dense point clouds and robust 3D landmarks in the map when a keyframe is captured. A dense depth map D_{deep} is derived from a pair of stereo images via the unsupervised deep learning-based optical flow [10]. Given the associated camera intrinsic and extrinsic parameters, the dense point cloud can be recovered from the dense depth map. KLT tracking [11] is used to identify good 2D features to subsequently track and estimate a sparse depth map D_{KLT} of the features from a pair of stereo images. KLT tracking tends to select features on relative well-textured regions, which ensures those features to be easily and accurately tracked on the following frames. However, the D_{KLT} usually contains outliers, and D_{deep} may be invalid in some regions, as the learning-based optical flow predicts on unseen images. We select robust 3D landmarks by comparing D_{KLT} and D_{deep}. If $|D_{KLT}(p) - D_{deep}(p)| < th_d$, where th_d is the threshold to filter outliers, the identified features, and associated depth values are selected as robust 3D landmarks.

2) Tracking: This module infers the current camera location relative to the previous camera position and updates the camera pose in the map. The 3D landmarks identified in the mapping module are projected onto the previous frame using the estimated camera pose of the previous frame and then tracked on the current frame via Lucas-Kanade (L-K) optical flow [11]. Given a set of 3D landmarks and their corresponding 2D projections, the current camera pose relative to the previous frame can be solved via the Perspective-n-Point (PnP) RANSAC, detailed in [2]. The pose is then transformed into the world coordinate system (the first frame in the sequence) and stored in the map.

3) Keyframe Insertion: The keyframe insertion module determines whether the current frame is a keyframe based upon the spatial distribution of tracked features in the current frame. Inserting a keyframe based on the number of tracked features [7] cannot guarantee that the tracked features are not over-concentrated in a small region of the current image due to camera motion. We grid the current frame into N equally sized, non-overlapping patches, then we insert the current frame as a keyframe if $N_p/N < th_p$, where N_p is the number of patches that contain tracked features, and th_p is a threshold. Our insertion rule ensures that there is enough area covered by tracked features, if not, the mapping module will detect new 3D landmarks from the current pair of stereo images. Also, this rule reduces the number of highly overlapping keyframes, thus decreasing computational burden, and improve reconstruction since the fusion of highly overlapping frames results in a blurry surface.

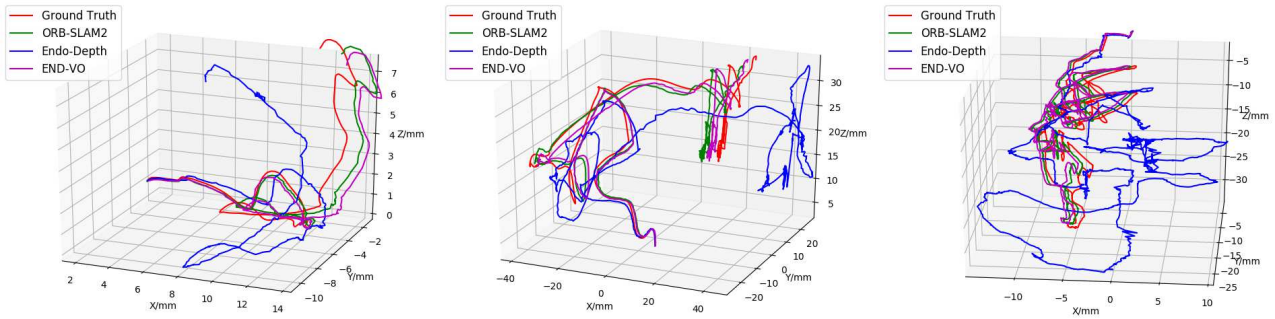
4) Local Bundle Adjustment: This module jointly optimizes the 3D landmarks and camera poses of the most recent N_k keyframes by minimizing re-projection error [7], leading to reduced tracking drift.

5) Visualization and Reconstruction: All camera poses, landmarks and point clouds of the map are visualized in

TABLE I

PERFORMANCE COMPARISON OF THREE VISUAL ODOMETRY / SCENE RECONSTRUCTION TECHNIQUES (END-VO - PROPOSED METHOD, ORB-SLAM2 AND ENDO-DEPTH) ON THE SCARED DATASET.

	ATE RMSE (mm)			**C2M RMSE (mm)**	
	ORB-SLAM2 [7]	Endo-Depth [3]	END-VO (Proposed Method)	Endo-Depth [3]	END-VO (Proposed Method)
Dataset1/video2	0.87	3.91	1.14	3.71	0.62
Dataset2/video2	4.63	21.50	3.85	1.98	0.37
Dataset3/video2	1.10	9.32	1.37	2.90	1.60



Datset1/video2 Datset2/video2 Datset3/video2

Fig. 2. Endoscopic camera trajectories estimated by ORB-SLAM2 [7], Endo-Depth [8], and our proposed END-VO on video sequences of the SCARED dataset. Note that the distance between the ground truth and END-VO trajectories is less than 4 mm across all reconstructed paths.

real time. We apply the Truncated Signed Distance Function (TSDF) [5] to the point clouds to reconstruct a global 3D mesh. The TSDF grids the space into equal voxels of the TSDF volume (V_{tsdf}) and sequentially averages the 3D locations and point cloud colors within each voxel.

B. END-VO Experiments, Implementation and Evaluation

Our experiments are conducted on three sub-datasets (dataset 1, 2, 3) of the SCARED dataset [12], collected from the abdominal anatomy of a porcine cadaver using a Da Vinci Xi surgical robot and a structured light. Each sub-dataset corresponds to a single porcine subject and contains 4 or 5 video sequences. Each video is accompanied by its associated camera calibration parameters, camera poses, and the point cloud of the first frame. The point clouds were reconstructed using structured light, while the camera poses were determined based on the robot kinematics. As we include the deep learning-based method in our framework, we use video 2 that is the longest video of each sub-dataset for testing and the rest of the available video sequences for training and validation. There are 5035 training pairs, 915 validation pairs, and 3997 testing pairs in all.

The proposed stereo visual odometry method was implemented in C++ and built upon the code library in [2]. We use the OpenCV implementation of KLT and PnP, and the Open3D implementation of TSDF. The unsupervised optical flow method was trained using the same parameter settings recommended by [10]. All experiments were conducted on a 2.60GHz Intel i7-9750H CPU and a GTX 2070 GPU. The hyper-parameters were set: $th_d = 8$, $N = 64$, $th_p = 0.65$, $N_k = 7$, $V_{tsdf} = 0.02\ mm^3$, tuned on training and validation dataset.

We evaluated our framework using two metrics, consistent with other reported methods [4], [8]. Endoscope tracking was evaluated based on the Root-Mean-Squared Absolute Trajectory Error (ATE-RMSE), computed as the root-mean-squared distance between the ground truth endoscope trajectory and the reconstructed trajectory, implemented in the evo Python library[1]. Scene reconstruction was evaluated using the Root-Mean-Squared Cloud-to-Mesh Error (C2M-RMSE), computed as the root-mean-squared error of the signed distance between the ground truth point cloud and the reconstructed mesh model, implemented using the open-souce CloudCompare tool. Note that the SCARED dataset only provides ground truth point cloud data for the first frame of all video sequences.

III. RESULTS

We assessed the performance of our proposed method (END-VO) against the ground truth data by comparing its performance to two other methods - ORB-SLAM2 [7] and Endo-Depth [3] - evaluated against the same ground truth data. The former (ORB-SLAM2) is a state-of-the-art sparse feature-based simultaneous localization and mapping (SLAM) system that includes global bundle adjustment in addition to visual odometry. The latter (Endo-Depth) is a recently published method that depends on dense depth maps estimated from a self-supervised depth estimation network

[1] `http://github.com/MichaelGrupp/evo`

trained on stereo images and photometric constraints for tracking.

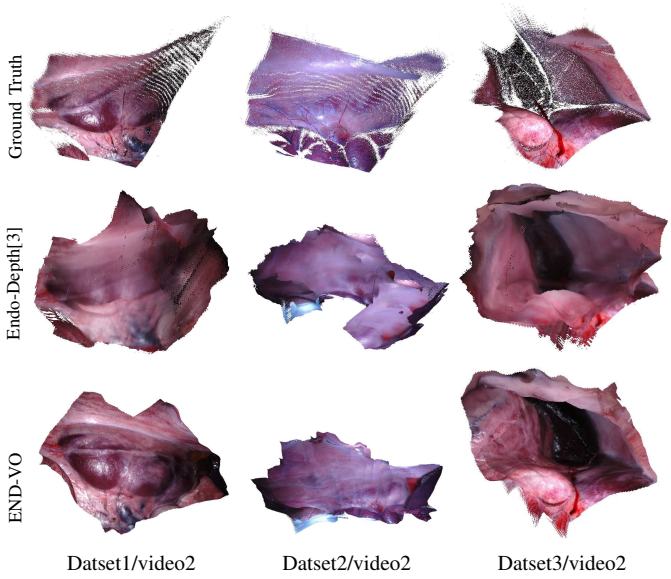


Fig. 3. Visualizations of scene reconstructions from the first frame of several video sequences of the SCARED dataset: ground-truth; Endo-depth [8]; and END-VO. Note the superior quality of the END-VO reconstructions relative to the blurry and incomplete Endo-depth reconstructions.

With regards to endoscope tracking accuracy, both the RMS-Absolute Trajectory Error and RMS-Cloud-to-Mesh Error reported in Table I and Fig. 2, respectively, confirm that END-VO performs comparably to ORB-SLAM2, while significantly out-performs Endo-depth. ORB-SLAM2's global bundle adjustment could be useful for tracking long image sequences in broad spaces, such as in autonomous driving. In abdominal cavities, local bundle adjustment may be sufficient. Endo-depth's performance lags both ORB-SLAM2 and END-VO, mainly because it does not include the bundle adjustment to reduce the accumulated camera drift, and photometric constraints are not commonly valid due to illumination change, which may lead to tracking errors. In terms of surgical scene reconstruction, ORB-SLAM2 is not capable to perform dense scene reconstructions. Both the RMS-Cloud-to-Mesh error reported in Table I and the qualitative reconstructions illustrated in Fig. 3 confirm that END-VO achieves a reconstruction quality superior to that of Endo-depth. Fig. 3 clearly shows that the surfaces reconstructed using Endo-depth are blurry, as a result of colors of overlapping point clouds being averaged during the reconstruction. This artifact is a result of Endo-depth selecting a keyframe after a fixed number of frames, causing significant point cloud overlap; on the other hand, END-VO mitigates such artifacts thanks to the keyframe insertion module that prevents significant point cloud overlap.

IV. Conclusions

We have presented a hybrid visual odometry framework (END-VO) for stereo endoscopy applications that enables both accurate endoscope tracking and dense scene reconstruction from stereo endoscopy. We used an unsupervised learning-based optical flow method to estimate dense depth maps from the low-textured tissue surfaces. We selected accurate and easy-to-track landmarks by leveraging the joint power of KLT tracking and unsupervised optical flow. We also designed an objective rule to govern keyframe insertion within our framework, which reduces computational burden and ensures high-quality scene reconstruction.

In summary, our proposed END-VO framework exploits the benefits of both traditional video odometry and unsupervised deep learning-based optical flow, therefore achieving high performance in both endoscope tracking and scene reconstruction from routine stereo endoscopy video sequences. Future work will focus on further improving and adapting this framework to enable accurate scene reconstruction from tissue-deforming surgical scenes.

References

[1] Angela Sorriento, Maria Bianca Porfido, Stefano Mazzoleni, Giuseppe Calvosa, Miria Tenucci, Gastone Ciuti, and Paolo Dario, "Optical and electromagnetic tracking systems for biomedical applications: A critical review on potentialities and limitations," *IEEE reviews in biomedical engineering*, vol. 13, pp. 212–232, 2019.

[2] Xiang Gao, Tao Zhang, Yi Liu, and Qinrui Yan, *14 Lectures on Visual SLAM: From Theory to Practice*, Publishing House of Electronics Industry, 2017.

[3] David Recasens, José Lamarca, José M Fácil, JMM Montiel, and Javier Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7225–7232, 2021.

[4] Ruibin Ma, Rui Wang, Yubo Zhang, Stephen Pizer, Sarah K McGill, Julian Rosenman, and Jan-Michael Frahm, "Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy," *Medical image analysis*, vol. 72, pp. 102100, 2021.

[5] Brian Curless and Marc Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[6] Jingwei Song, Jun Wang, Liang Zhao, Shoudong Huang, and Gamini Dissanayake, "Mis-slam: Real-time large-scale dense deformable slam system in minimal invasive surgery based on heterogeneous computing," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 4068–4075, 2018.

[7] Raul Mur-Artal and Juan D Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[8] Kutsev Bengisu Ozyoruk, Guliz Irem Gokceler, Taylor L Bobrow, Gulfize Coskun, Kagan Incetan, Yasin Almalioglu, Faisal Mahmood, Eva Curto, Luis Perdigoto, Marina Oliveira, et al., "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Medical image analysis*, vol. 71, pp. 102058, 2021.

[9] Amir Atapour-Abarghouei and Toby P Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2800–2810.

[10] Zixin Yang, Richard Simon, Yangming Li, and Cristian A Linte, "Dense depth estimation from stereo endoscopy videos using unsupervised optical flow methods," in *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2021, pp. 337–349.

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

[12] Max Allan, Jonathan Mcleod, Congcong Wang, Jean Claude Rosenthal, Zhenglei Hu, Niklas Gard, Peter Eisert, Ke Xue Fu, Trevor Zeffiro, Wenyao Xia, et al., "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.