Topological SLAM in colonoscopies leveraging deep features and topological priors

Javier Morlana, Juan D. Tardós and José M. M. Montiel

i3A, University of Zaragoza, Spain
{jmorlana,tardos,josemari}@unizar.es

Abstract. We introduce ColonSLAM, a system that combines classical multiple-map metric SLAM with deep features and topological priors to create topological maps of the whole colon. The SLAM pipeline by itself is able to create disconnected individual metric submaps representing locations from short video subsections of the colon, but is not able to merge covisible submaps due to deformations and the limited performance of the SIFT descriptor in the medical domain. ColonSLAM is guided by topological priors and combines a deep localization network trained to distinguish if two images come from the same place or not and the soft verification of a transformer-based matching network, being able to relate far-in-time submaps during an exploration, grouping them in nodes imaging the same colon place, building more complex maps than any other approach in the literature. We demonstrate our approach in the Endomapper dataset, showing its potential for producing maps of the whole colon in real human explorations. Code and models are available at: github.com/endomapper/ColonSLAM

Keywords: Topological SLAM · Deep features · Colonoscopy

1 Introduction

The interest for medical Computer Vision has been growing in the last decades, with many works being able to successfully apply classical and modern techniques to the medical domain. In this context, Simultaneous Localization And Mapping (SLAM) is a line of research that has been receiving huge attention due to its broad spectrum of possible applications such as medical robotics and navigation assistance. SLAM systems aim to localize a camera while building a map of an unexplored environment. Two kinds of representations can be obtained by SLAM algorithms. Metric SLAM estimates a 6DoF camera trajectory and a geometric 3D point cloud, while topological SLAM obtains a graph whose nodes represent places that can be connected by covisibility or traversability.

We are interested in the colonoscopy domain, a field of medicine in which technology still has little presence. Typically, practitioners manoeuvre through the colon anatomy based on prior knowledge and experience, visualizing the raw endoscopy images in the screen without any other information input. However, metric SLAM struggles in colonoscopies due to illumination changes that hinder

keyframe registration, and dynamic elements and deformations that violate the rigidity constraint. The result are small and disconnected 3D submaps, quite different from the long maps obtained in out-of-the-body scenes.

We propose ColonSLAM, a topological SLAM system for the colonoscopy domain, where nodes are groups of small metric submaps imaging the same colon place. We build on top of a recent metric SLAM that builds small submaps using classical image features, and we continuosly find relationships between far-in-time submaps, leveraging deep global visual place recognition descriptors, transformer-based matching techniques and topological connectivity priors. Our contributions in this work are threefold:

- We present ColonSLAM, the first metric-topological SLAM system able to map the whole colon creating a graph that codes the procedure complexity.
- We propose a novel visual place recognition network $\mathbb{L}$, able to identify covisible images to build a topological map from submaps obtained by metric SLAM.
- We perform an evaluation in real human colonoscopy data, showing our ability to build complex maps to cover the entire colon exploration.

2 Related Work

Metric SLAM solutions already work well in natural scenes, being able to map unknown environments through feature-based approaches [22,7], which employ geometric bundle adjustment, or direct methods such as [12,11], optimizing errors in the photometric space. Nowadays, there is a growing interest in bringing SLAM to the medical domain. Mahmoud et al [19] applies ORB-SLAM [22] to laparoscopy, SAGE-SLAM [15] integrates learned depth and features to reconstruct endonasal surgery scenes, and RNN-SLAM [18] combines DSO [11] with learned depth to create dense reconstructions of the colon. The recent approach CudaSIFT-SLAM [10] builds on the ORB-SLAM3 multi-mapping system [7] replacing ORB features by CudaSIFT [6], building metric multi-maps in human colon in real-time. It produces small disjoint 3D maps, where covisibility between the keyframes in each map is guaranteed as every keyframe goes through several stages of filtering: matching, geometric verification, 3D triangulation and geometric bundle adjustment. Multi-maps are key for robustly dealing with tracking losses due to occlusions, deformation and motion blur prevalent in colonoscopy.

RNN-SLAM and CudaSIFT-SLAM are currently the top performers in colonoscopic SLAM, but they are unable to relate far-in-time submaps representing the same place. We build on the output of CudaSIFT-SLAM to obtain meaningful topological maps by establishing relationships between their disjoint submaps.

Topological SLAM avoids the geometry estimation and focuses on aggregating covisible images by their appearance, leveraging on visual place recognition (VPR) methods. These algorithms can be better suited for the medical domain, where metric SLAM tends to fail due to deformations or occlusions. Classical methods [9,1,13] converted local features such as SIFT [16] or ORB [25] into a

Bag-of-Words representation, finding the most similar images, further verified by geometry in order to close a loop between nodes. Recently, ColonMapper [21] leveraged the Bayesian filtering proposed in [2,1] with global deep features for VPR to build topological maps with a trivial two-node connectivity which links each node with its anterior and posterior in time neighbours. Despite its simplicity, ColonMapper is able to map the whole colon, and remarkably, the map was reused for topological localization two weeks afterwards, in a second colonoscopy of the same patient. While ColonMapper builds the map and afterwards localizes, our ColonSLAM performs a proper topological SLAM, simultaneously localizing and updating the map in the processing of each new incoming submap.

Our proposal is also close to recent works building topological graphs with the help of deep learning [8,23,26]. They build a graph using retrieval networks as in [21], but tailoring it as means to an end, focusing on robot navigation or affordances learning. Differently from them, we focus on building the graph that defines the topological map, as creating meaningful representations is not straightforward in the medical domain. Colonoscopy images, in particular, are a challenging task for visual recognition algorithms due to their weak texture and the visual similarity of different regions.

Neural Networks for Visual Place Recognition are also closely related to our work. The works of [20,17,21] brought popular image retrieval networks [24,3] to the colonoscopy domain, and a similar approach was followed for topological graphs in out-of-the-body scenes in [8,23,26]. ColonMapper [21], particularly, employed an image retrieval network trained by means of a margin loss and used it to for trivial topological map creation and its posterior localization. We empirically found that this training objective is not discriminative enough to build non-trivial maps with high precision. Our work also leverages on the transformer-based matcher LightGlue [14] to establish relationships between consecutive nodes, while ColonMapper tried a similar strategy with LoFTR [28].

3 ColonSLAM

3.1 Node building

We create a topological map $G = (N, E)$ composed by nodes N and edges E. Each node represents a *place*, a distinctive section of the colon, while edges link traversable places connected in space. ColonMapper [21] assumed a simplistic graph of consecutive places connected with its two closest neighbours computed from a global descriptor similarity and a matching verification with LoFTR. In contrast, we propose to sequentially build a full-fledged topological map which captures the complex covisibility and traversability among the metric submaps.

The starting point of the topological map in ColonSLAM is a set of metric 3D disjoint *submaps* obtained by CudaSIFT-SLAM with their linear connectivity, i.e. each submap is connected only with its anterior and posterior submaps (see Fig. 1). They are composed of several keyframes (distinctive images). For each keyframe, we extract a global descriptor $\mathbf{d} \in \mathbb{R}^D$ by means of a localization network $\mathbb{L}$ (Sec. 3.2). The submaps obtained are usually small, typically 5

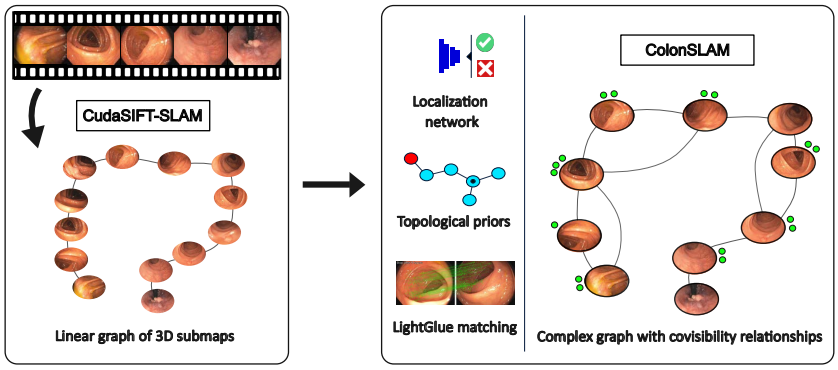

Fig. 1. ColonSLAM. From a linear graph of metric submaps, ColonSLAM is able to obtain a topological graph with rich connections by leveraging a novel localization network, topological priors and LightGlue matching.

seconds lifespan and 15 keyframes. Some of the maps are taken from the same colon location, but CudaSIFT-SLAM was not able to merge them together. As our objective is to build rich graphs for the colon, we use a similar formulation as [23], considering our node as a colon region that encompasses several submaps observing that particular region. For example, if CudaSIFT-SLAM reconstructed the submaps s_{20}, s_{22} and s_{23} for the cecum area, our graph would represent the cecum as a node $n \in N$ with submaps $\{s_{20}, s_{22}, s_{23}\}$. We discard *in-between* frames, chunks of video between submaps that were not included in any CudaSIFT-SLAM 3D model, as they are generally noisy observations containing unmappable frames, i.e. blurry, occluded or covered by fluids.

3.2 Localization Network

Our localization network $\mathbb{L}$ predicts if two images come from the same place or not, and we use it to determine if the incoming *submap* is already included in the map. The network is composed of a backbone and a 5-layer MLP. The backbone is initialized from the endoscopy foundational model EndoFM [29], which, for an image I extracts a global descriptor $\mathbf{d} \in \mathbb{R}^{768}$. To decide if two images I_A, I_B come from the same place, we subtract their descriptors $g = d_A - d_B$ and feed g to the MLP followed by a softmax, predicting a similarity score sim that allows to decide if they come from the same place, as can be observed in Fig. 2. We fine-tune the last two layers of the backbone and the MLP using a cross-entropy objective. Training details are explained in Sec. 4.1.

3.3 Topological Simultaneous Localization and Mapping

ColonSLAM receives a linear topological graph formed by all submaps from CudaSIFT-SLAM. The main idea of ColonSLAM is to identify which submaps

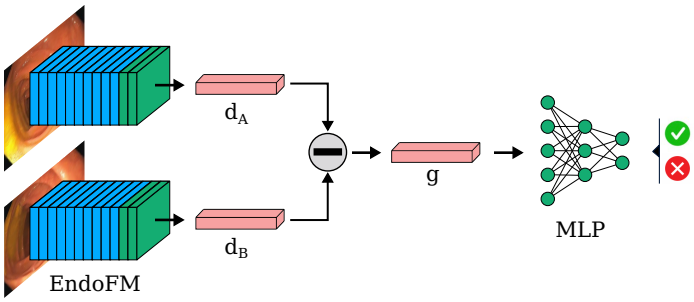

Fig. 2. Localization network $\mathbb{L}$. It obtains a *sim* score, deciding if two images come from the same place. The backbone green blocks and the MLP are fine-tuned.

represent the same colon location, merging them into the same node. This observation capability builds traversability links between distant nodes, resulting in a richer graph than the linear one. Previous work [21] followed a Bayesian approach for localizing a new exploration of a patient against a trivial map of this patient built in a previous exploration. We empirically found that using a Bayesian approach to simultaneously map and localize decreases the performance of topological SLAM. For this reason, we opted for a simpler yet effective approach that allows us to leverage on topological priors without explicitly modelling the localization probability. We demonstrate that these priors are helpful for all methods, specially in combination with the discriminating power of our localization network $\mathbb{L}$ and the matching capabilities of LightGlue [14].

Topological priors. Our topological SLAM starts with the first *submap* s_0, that initializes the first node N_0. For every submap s_i, we extract a descriptor d_k with $\mathbb{L}$ for every keyframe $I_k \in s_i$. Here we leverage on colonoscopy priors, which are linear by nature. In a typical colonoscopy, the practitioner reaches the cecum as fast as possible, and then, performs a slow exploration in the withdrawal stage. Occasionally, the camera moves back and forth, i.e. when the endoscope is obstructed or the practitioner is exploring carefully a particular area. In any case, the camera has to observe close-by areas to the current location before reaching further places. It is physically impossible to go from the cecum to the transverse without going through the ascending colon first. For this reason, we establish our *topological connectivity prior* as a search space ω where a covisible node can be looked for. The search space ω is defined as a window of nodes at a distance equal or smaller than m nodes from the previous position S_{t-1}.

Selecting a localization S_t**.** We compute the score *sim* of $I_k \in s_i$ against every node $n \in \omega$. We compare d_k against all the image descriptors in each $s_j \in n$ with our localization network $\mathbb{L}$. The score for each submap s_j is equal to the average of the top-3 ranked images from s_j, while the score for the node n is the highest score among its submaps. Besides, for every keyframe I_k from s_i we store in l_{sim} the node n_j with the highest score and its value. We define the $score_{\mathbb{L}}$ as the median value of the scores for the node n_j that with higher occurences in l_{sim}.

We are interested in determining if the incoming submap s_i belongs to a node n_j in the graph or not, in order to add s_i to n_j or create a new node n_{new} connected to the previous position S_{t-1}. We have two ways of triggering a localization: a) LightGlue finds $m_{LG} > th_{LG}$ matches between any image in s_i and any image from any node $n_j \in \omega$ (Eq. 1) or b) $score_{\mathbb{L}} > th_{sim}$ (Eq. 2):

$$m_{LG} > th_{LG} \tag{1}$$

$$score_{\mathbb{L}} > th_{sim} \tag{2}$$

The reasoning behind bypassing LightGlue in condition b) is that, despite its matching abilities and precision, LightGlue is not able to deal with all the challenges in colonoscopies, failing when images are far from each other. $\mathbb{L}$ is able to reliably find some of this cases, so we chose to complement the two methods, looking for higher recall values while keeping an acceptable precision ($\sim 90\%$). If a localization is accepted in n_j, the current position S_t is set to n_j, otherwise is set to n_{new}, adding a traversability link with previous position S_{t-1}.

4 Experiments

4.1 Implementation details

Localization network $\mathbb{L}$ training. We train our localization network $\mathbb{L}$ with the Endomapper [4] training data proposed in [21]. We use the already labelled data to extract samples. Labels were obtained in [21] using COLMAP [27] and manual labelling. It includes positive examples from COLMAP and some hard positives manually labelled, besides per-cluster covisibility labelling that allows extracting negative pairs from the same sequence. For our cross-entropy loss, we train with pairs query-positive and query-negative, trying to predict if the images are similar or dissimilar, respectively. We get one random positive sample for each query from the positives pool, while we always provide the hardest negative coming from the same sequence as the query, based on the global descriptor distance. [23] trained with concatenated vectors, while we found crucial for our network's convergence to subtract them before passing the result to the MLP. We fine-tune the last two layers of EndoFM [29] and the MLP, freezing the rest of the network. Convergence is achieved after 4 epochs based on the cross-entropy loss in the validation set, using 10k queries per epoch and re-mining negatives every 2500 queries. Our training framework is based on [5].
Other details. We use off-the-shelf LightGlue [14], reducing the SuperPoint detection threshold and disabling early stoppers from LightGlue in order to get the most reliable matches. The matching acceptance threshold is $th_{LG} = 100$.

4.2 Evaluation on the Endomapper dataset

We selected two sequences of the Endomapper dataset as our ground truth. We chose the same sequences as ColonMapper (Seq_027 and Seq_035), the

Method	Seq_027		Seq_035		Average		Runtime
	Precision	Recall	Precision	Recall	Precision	Recall	
Morlana21 [20]	0.83	0.51	0.88	0.49	0.85	0.50	38 s
+ Topologic prior	0.88	0.50	<u>0.95</u>	0.47	0.91	0.48	36 s
R50-NV-H [21]	0.64	0.42	**0.97**	0.37	0.80	0.39	50 s
+ Topologic prior	0.78	0.45	**0.97**	0.33	0.87	0.39	39 s
LightGlue [14]	**1.0**	0.45	0.91	0.33	<u>0.95</u>	0.39	∼56 min
+ Topologic prior	**1.0**	0.45	0.94	0.32	**0.97**	0.38	∼11 min
$\mathbb{L}$ (ours)	0.87	<u>0.64</u>	0.76	<u>0.68</u>	0.81	<u>0.66</u>	∼1 min
+ Topologic prior	<u>0.96</u>	0.61	0.92	0.67	0.94	0.64	50 s
+ LightGlue	0.94	**0.70**	0.87	**0.70**	0.90	**0.70**	∼25 min

Table 1. Precision and Recall results. Bold: best. Underlined: second best.

closest work to ours, easing the comparison. Labeling was done following the text footage available in the Endomapper dataset, created by the doctor during the exploration. We first process them with CudaSIFT-SLAM, obtaining a set of *submaps* $\in \{s_0, \ldots, s_n\}$. We manually labelled which submaps are covisible, that is, should belong to the same node. Two nodes are covisible if they observe the same location. We labelled both medium-covisible relationships and long-term covisibility, i.e. a polyp seen both in the entry and the withdrawal phase. Additionally, *submaps* are labelled chronologically: we know if the incoming *submap* should be localized against previous nodes or if it should create a new node.

We show precision and recall values in Table 1. We compare against related methods to our work: Morlana21 [20] and R50-NV-H (from ColonMapper) [21], two image retrieval networks trained for the colonoscopy domain, and LightGlue [14], a state-of-the-art network in image matching, with enough matching power to establish correspondences between close nodes in colonoscopies. ColonMapper also proposed a localization algorithm where mapping was not considered, so its application here is not straightforward. Instead, we evaluate the network proposed in their work. Besides, we provide an ablation study of the three main elements of our pipeline: the localization network $\mathbb{L}$, the addition of the topological prior and LightGlue matching. Precision and recall are defined as:

$$\mathbf{P} = \frac{TP}{TP + FP} \quad , \quad \mathbf{R} = \frac{TP}{TP + FN} \tag{3}$$

TP are true positives, correctly localized *submaps*. A localization for *submap* s_i for node n_j is deemed correct if the majority of *submaps* in n_j were labelled as positives with s_i. FP are false positives, wrongly localized *submaps*. FN are false negatives, *submaps* that should be localized but instead started a new node.

We evaluate the performance of the different methods and the benefits of the topological prior. For retrieval networks (Morlana21 [20], R50-NV-H [21] and $\mathbb{L}$), we accept a localization only if Eq. 2 is fulfilled. We apply a different

threshold for each of the networks as the score distribution given by each network is different. To allow a fair comparison between them, we tuned the best threshold for every network in terms of precision-recall performance. For Morlana21 [20], $th_{sim} = 0.85$; for ColonMapper [21], $th_{sim} = 0.65$ and for $\mathbb{L}$, $th_{sim} = 0.95$. For LightGlue [14], we compare first, medium and last image in s_i against the first, medium and last image of all nodes, as comparing all is too expensive. If any comparison fulfills Eq. 1, we stop the computation and accept the localization.

Approaches without the topological prior search along the whole graph, while when the topological prior is added, the search is only allowed in the window ω, with $m = 5$. For our full approach ($\mathbb{L}$ + Topological prior + LightGlue), we accept a localization if Eq. 1 or 2 are fulfilled as explained in Sec. 3.3. All approaches improve their precision significantly when the topological prior is applied, specially for our network $\mathbb{L}$, that receives a great boost in precision while getting the highest recall. Reducing the search space using the topological graph information is helpful for image retrieval networks, as they are not confused by similar frames coming from far regions. The effect in LightGlue is minimal, as it is only able to match close-by images, but it reduces computation time by 5x while maintaining the performance. Our network $\mathbb{L}$, in combination with the topological prior, is able to compete with LightGlue precision while getting an improvement of +70% in recall an being several orders of magnitude faster. Despite this, we aim to find at most connections as possible (high recall) while having a reasonable precision. When complementing our approach with LightGlue, we finally obtain a precision of 90% with a recall of 70%.

In Figure 3 we show a comparison between the CudaSIFT-SLAM graph and the result of our approach. Green and red dots represent correctly and wrongly localized submaps within a node, respectively. As it can be seen, we are able to build a complex graph with dozens of submaps correctly localized. The traversability connections faithfully show how the exploration was made: quickly during the entrance until the cecum was reached, showed with few traversability links, and then some exploration and back and forth movements, represented as a lot of traversability edges in the ascending colon.

5 Conclusions

We have presented ColonSLAM, the first topological SLAM able to build rich graphs of the whole colon, capturing the complexity of the colonoscopy exploration. Leveraging on our robust localization network and guided by topological priors, ColonSLAM is able to reliably build a graph by finding traversability and covisibility connections between distant nodes. The graphs obtained with ColonSLAM will serve as personalized patient maps, paving the way to assisted navigation and disease monitoring in colonoscopy. In future work, we will focus on finding even longer term relationships i.e. entry-withdrawal and second explorations of the same patient as they are a limitation for ColonSLAM. Finding these long-term correspondences is the key to the building and exploitation of personalized patient maps.

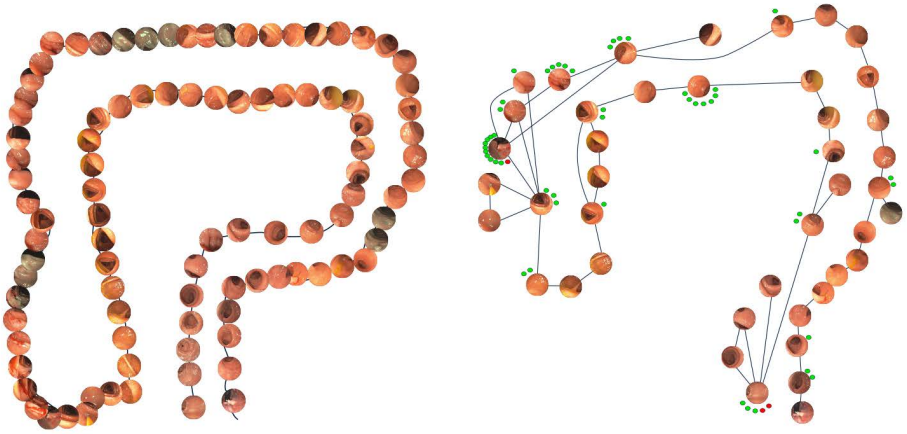

Fig. 3. Seq_027 topological map. CudaSIFT-SLAM (left) and ColonSLAM (right).

Acknowledgments. Work supported by EU-H2020 grant 863146: ENDOMAPPER, Spanish grant PID2021-127685NB-I00, Aragón grant DGA_T45-17R.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angeli, A., Doncieux, S., Meyer, J.A., Filliat, D.: Incremental vision-based topological slam. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 1031–1036 (2008)
2. Angeli, A., Filliat, D., Doncieux, S., Meyer, J.A.: Fast and incremental method for loop-closure detection using bags of visual words. IEEE Transactions on Robotics **24**(5), 1027–1037 (2008)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5297–5307 (2016)
4. Azagra, P., Sostres, C., Ferrández, Á., Riazuelo, L., Tomasini, C., Barbed, O.L., Morlana, J., Recasens, D., Batlle, V.M., Gómez-Rodríguez, J.J., et al.: Endomapper dataset of complete calibrated endoscopy procedures. Scientific Data **10**(1), 671 (2023)
5. Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., Caputo, B.: Deep visual geo-localization benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5396–5407 (2022)
6. Björkman, M.: CudaSIFT. `https://github.com/Celebrandil/CudaSift` (2007), [Online; accessed 05-April-2023]
7. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM. IEEE Transactions on Robotics **37**(6), 1874–1890 (2021)

8. Chaplot, D.S., Salakhutdinov, R., Gupta, A., Gupta, S.: Neural topological slam for visual navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12875–12884 (2020)
9. Cummins, M., Newman, P.: FAB-MAP: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research **27**(6), 647–665 (2008)
10. Elvira, R., Tardós, J.D., Montiel, J.M.: CudaSIFT-SLAM: multiple-map visual SLAM for full procedure mapping in real human endoscopy. arXiv preprint arXiv:2405.16932 (2024)
11. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2017)
12. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. pp. 834–849 (2014)
13. Galvez-Lopez, D., Tardos, J.D.: Real-time loop detection with bags of binary words. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 51–58 (2011)
14. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: International Conference on Computer Vision (2023)
15. Liu, X., Li, Z., Ishii, M., Hager, G.D., Taylor, R.H., Unberath, M.: SAGE: SLAM with appearance and geometry prior for endoscopy. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 5587–5593 (2022)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**, 91–110 (2004)
17. Ma, R., McGill, S.K., Wang, R., Rosenman, J., Frahm, J.M., Zhang, Y., Pizer, S.: Colon10k: a benchmark for place recognition in colonoscopy. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1279–1283 (2021)
18. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Medical image analysis **72**, 102100 (2021)
19. Mahmoud, N., Collins, T., Hostettler, A., Soler, L., Doignon, C., Montiel, J.M.M.: Live tracking and dense reconstruction for handheld monocular endoscopy. IEEE Transactions on Medical Imaging **38**(1), 79–89 (2019)
20. Morlana, J., Azagra, P., Civera, J., Montiel, J.M.: Self-supervised visual place recognition for colonoscopy sequences. In: Medical Imaging with Deep Learning (MIDL) (July 2021)
21. Morlana, J., Tardós, J.D., Montiel, J.M.M.: ColonMapper: topological mapping and localization for colonoscopy. In: IEEE Int. Conf. Robotics and Automation (2024)
22. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
23. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 163–172 (2020)
24. Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE transactions on pattern analysis and machine intelligence **41**(7), 1655–1668 (2018)
25. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: An efficient alternative to SIFT or SURF. In: 2011 International conference on computer vision. pp. 2564–2571. Ieee (2011)

26. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. In: International Conference on Learning Representations (2018)
27. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
28. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. CVPR (2021)
29. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)