Real-time lumen detection for autonomous colonoscopy

Al-Bander, Baidaa; Mathew, Alwyn; Magerand, Ludovic; Trucco, Manuel; Manfredi, Luigi

Published in:
Imaging Systems for GI Endoscopy, and Graphs in Biomedical Image Analysis

DOI:
10.1007/978-3-031-21083-9_4

Publication date:
2022

Document Version
Peer reviewed version

Real-Time Lumen Detection for Autonomous Colonoscopy

Baidaa Al-Bander[1], Alwyn Mathew[1], Ludovic Magerand[2], Emanuele Trucco[2], and Luigi Manfredi[1,⋆]

[1] School of Medicine, University of Dundee, Dundee, UK
l.manfredi@dundee.ac.uk
[2] School of Science and Engineering, University of Dundee, Dundee, UK

Abstract. Lumen detection and tracking in the large bowel is a key prerequisite step for autonomous navigation of endorobots for colonoscopy. Attempts at detecting and tracking the lumen so far have been made using optical flow and shape-from-shading techniques. In general, these methods are computationally expensive, and most are either not real-time nor tested on real devices. To this end, we present a deep learning-based approach for lumen localisation from colonoscopy videos. We avoid the need for extensive, costly annotations with a semi-supervised learning and a self-training scheme, whereby only a small subset of video frames is annotated. We develop an end-to-end pseudo-labelling semi-supervised approach incorporating a self-training scheme for colon lumen detection. Our approach reveals a competitive performance to the supervised baseline model with both objective and subjective evaluation metrics, while saving heavy labelling costs in terms of clinicians' time. Our method for lumen detection runs at $60ms$ per frame during the inference phase. Our experiments demonstrate the potential of our system in real-time environments, which contributes towards improving the automation of robotics colonoscopy.

Keywords: Autonomous colonoscopy · semi-supervised learning · lumen detection · self-training · endorobots for colonoscopy · bowel cancer.

1 Introduction

Colorectal cancer (CRC) is the third cause of cancer-related mortality worldwide, after lung and breast cancer [1]. Colonoscopy is regarded as the main clinical diagnostic technique for CRC, with regular screening being a significant step in drastically reducing mortality rates. Optical colonoscopy (OC) is the gold standard for optical screening and treatment of CRC since it enables biopsy, pathological prediction and treatment [5]. However, the current generation of colonoscopes has limitations, such as patient pain and discomfort, narrow field of view, difficulties to detect lesions located behind colonic folds, time-consuming

⋆ Corresponding author

and complex procedure to learn [16]. Thus, developing low-risk, cost-effective, and more efficient alternative solutions for colonoscopy is now necessary. Rapid advancements in endorobotics have produced a new generation of systems that have the potential to overcome the above limitations. For instance, real-time visual feedback from a monocular camera can now be incorporated into the control loop to detect the region of haustral folds in the colon and determine the centre of the lumen [14, 15]. The deformable nature of the large bowel poses sensing and navigation challenges untackled by traditional robotics. Current localisation and navigation strategies for colonoscopy [8] generally depend on external hardware (i.e. permanent on-board magnet linked to an external magnetic field). Computer vision-based navigation and localisation, relying on feature recognition, can offer a solution, but, the deformable nature of the environment may cause significant difficulties to traditional feature location methods [21].

Several approaches to designing autonomous visual navigation systems for endoscopes using images have been reported [24]. Many are unsuitable for real-time operation or fail to work when the lumen centre is hard to detect. Despite these challenges, methods based on optical flow [11], shape from shading [6], structure from motion [10] and segmentation [17] have been developed for automatic navigation. The considerable variety in lumen feature appearance due to the surfaces in view, lighting and acquisition techniques makes it challenging to construct a universal model performing optimally in any environment and condition. In addition, further factors like occlusion, deformation, off-centre lumen can degrade performance. Deep learning (DL) algorithms offer great potential in medical image analysis and interpretation, supported by rapid improvements in GPU hardware. Endoscopists' performance in the diagnosis of adenoma or polyp [4] has also been shown to benefit from the assistance of deep learning systems. Ahmad et al. [2] reported a comprehensive review of studies that exploited artificial intelligence, especially DL models, in colonoscopy computer-aided diagnosis. Methods based on supervised learning (SL) typically require large quantities of labelled data annotated by experts to achieve high diagnostic accuracy. However, in the medical domain, only a limited amount of labelled data and a considerably greater amount of unlabelled data is available. Contrary to SL, semi-supervised learning (SSL) leverages both labelled and unlabelled data to offer a low-cost alternative to the time-consuming massive data labelling task [12, 20, 23, 25, 27, 28].

Our work develops a vision-based system harnessing deep neural networks to detect and track the lumen in real time, enabling reliable endorobot navigation colonoscopy. Unlike existing lumen detection models developed to work on specific video data, our model is developed to accommodate video data captured from a variety of environments,including synthetic, plastic phantom, and real colonoscopy videos. We introduce a fast and accurate method that controls the level of supervision needed, leveraging a semi-supervised scheme for lumen localisation. By exploiting a few labelled frames and a large number of unlabelled frames, we develop an end-to-end pseudo-labelling semi-supervised approach incorporating a self-training scheme for colon lumen detection. To evaluate ro-

bustness and reliability, we have conducted experiments on comprehensive video data. Results show promising recall and precision on lumen detection with plastic phantom and simulated datasets and suggest an excellent generalisation ability on unseen real colonoscopy videos. The results also demonstrate the benefits of the SSL strategy over the fully supervised scheme (baseline model), without sacrificing the run-time advantage or prediction accuracy.

2 Methods

Inspired by Xu et al. [27], who achieved competitive detection performance on natural image data, we propose to use a semi-supervised learning (SSL) scheme incorporating a self-training framework for colon lumen detection, as shown in Fig. 1. Our method exploits the mentor-student approach for a hybrid learning strategy. Both mentor and student have the same architecture, the default Faster R-CNN object detector model. First, the object detector model is trained with a classical supervised scheme from 40% of the labelled data, of which 2% are used for validation and hyper-parameters tuning. The object detector is then used as a mentor to generate pseudo-labels, as a test-time inference, from 30% of unlabelled data. The student is trained by mixing those pseudo-labelled data with an additional 10% of the labelled data, which is augmented. The remaining 20% of labelled data are used for testing.

Baseline Supervised object Detector Model. A single-level feature detector, Faster Region Based Convolutional Neural Networks (Faster R-CNN) [19] is

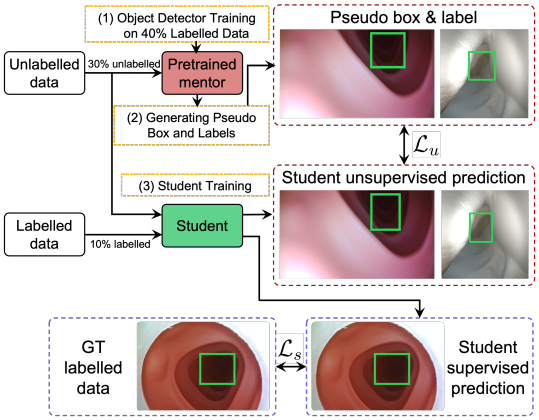

Fig. 1: Overview of mentor-student model training for lumen detection. Pseudo-labels (bounding boxes and class labels) are generated from a pre-trained mentor model with unlabelled data. Student unsupervised loss is computed with the pseudo-labels above a specific threshold in a semi-supervised manner. 10% of the labelled data with augmentation is also used to train the student model. GT: ground truth.

harnessed to produce the baseline model used as mentor in our SSL scheme. We trained it with a classical supervised scheme using 40% of the manually labelled frames. Faster R-CNN has two heads, one for object classification and the other for bounding boxes regression. It also has a fully convolutional Region Proposal Network (RPN) that takes the input features of frame and produces region proposals with an objectness score (denoting the probability of object or not object (background) for each proposal). The RPN predicts the offsets of region proposals from established reference boxes, known as anchor boxes. Anchor boxes are predetermined and fixed-size boxes distributed over the input frame with a variety of sizes and aspect ratios. A non-maximum suppression (NMS) algorithm [9] is then applied to filter out the predicted region proposals, depending on a confidence threshold score, which is set to value of 0.7. The advantage of employing box predictions after NMS over raw predictions (before applying NMS) is that it avoids duplicated and overlapped results. Once the region proposals are selected, the lumen object classification and boundary box regression are then measured in a supervised fashion. The supervised loss function used to learn the baseline model is:

$$\mathcal{L}_s = \frac{1}{N_l} \sum_{i=1}^{N_l} \left(\mathcal{L}_{\text{cls}} \left(p_i, p_i^* \right) + \mathcal{L}_{\text{reg}} \left(t_i, t_i^* \right) \right) \tag{1}$$

Where i indexes a labelled frame, p_i: predicted probability of proposal contains a lumen object or not, p_i^*: the ground-truth value of proposal contains a lumen object or not, t_i is the coordinates of the predicted lumen proposal, t_i^* is the ground-truth coordinate associated with the bounding box of the lumen, $\mathcal{L}_{\text{cls}}$ is the classification loss, $\mathcal{L}_{\text{reg}}$ is the bounding box regression loss, and N_l denotes the number of labelled frames in batch.

Semi-Supervised with Self-training Model. In our mentor-student learning scheme, the student is trained in a semi-supervised fashion integrating a self-training strategy which has achieved considerable success including Noise-Student [26], STAC [23], and SoftTeacher [27]. The phases of our SSL incorporated with self-training are:

1. Leverage the baseline pre-trained supervised detector model as a mentor model to generate pseudo-labels and pseudo-bounding box annotations for 30% of unlabelled frames. This process includes a forward pass of the Res50 backbone model, RPN and classification network, followed by the NMS post-processing. These predicted pseudo-labels and pseudo-bounding box annotations are considered the ground truth to compare with the prediction from the student model in an unsupervised loss function.
2. Train the student model using both those pseudo-labelled frames and 10% of the manually labelled data not seen by the mentor on which data augmentation is applied. This requires establishing a loss function for the student that sums the losses of both supervised and unsupervised models.

To compute the loss for pseudo-labelled frames when training the student, the generated pseudo-labels are used as ground-truth to be compared to the student

prediction, producing an unsupervised loss function as follows:

$$\mathcal{L}_u = \frac{1}{N_u} \sum_{i=1}^{N_u} \left(\mathcal{L}_{\text{cls}} \left(r_i, r_i^* \right) + \mathcal{L}_{\text{reg}} \left(s_i, s_i^* \right) \right) \tag{2}$$

Here i indexes an unlabelled frame, u an unlabelled frame, r_i refers to the predicted probability of proposal containing a lumen object, r_i^* represents the generated pseudo-label of proposal, s_i is the coordinates of predicted proposal for lumen, s_i^* is the pseudo-boxes of lumen generated by the mentor, N_u denotes the number of unlabelled frames. For the 10% labelled frames, the student calculates the loss between the provided ground truth and the predicted labels, via a supervised loss. The total loss of the student model is the weighted sum of the unsupervised and supervised losses, i.e., using Eq. (1) and Eq.(2):

$$\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_u \tag{3}$$

Where α denotes the weight of unsupervised loss, determined experimentally.

Inference and Refinement. Once the model is trained, it can be adopted for the inference phase to produce the predictions of bounding boxes over the lumen area on a frame basis. To maintain the temporal consistency among predicted bounding boxes on a sequence of consecutive frames, we carefully designed a simple yet effective strategy to choose the bounding box proposal that preserves a minimum distance to the bounding box in the previous frame in a sequence of frames. The application of refinement scheme assumes that the intersection over the union between the bounding boxes of two consecutive frames is not null, which typically results from an abrupt camera movement. This scheme is applied by locating the centre points $(x_{i,c}, y_{i,c})$ of the predicted bounding boxes in frame i, where $c \geq 1$. The centre points are computed from the predicted bounding boxes, represented by the value of top left corner (x_{min}, y_{min}) and bottom right corner (x_{max}, y_{max}). The centre point (x_c, y_c) in frame i is defined as follows:

$$(x_c, y_c) = (round(x_{min} + \frac{x_{max} - x_{min}}{2}), round(y_{min} + \frac{y_{max} - y_{min}}{2})) \tag{4}$$

Euclidean distance is measured among the centre points in frame i and the centre point in the previous frame, $i-1$. The centre point that achieves the minimum distance is then selected, and the bounding box accompanied by this point is produced as the outcome of the model prediction.

3 Experimental Set-up, Results and Discussion

Datasets. Public synthetic dataset [18] consisting of 16,016 RGB frames generated from the video is used in our study. The size of frames is 256×256 pixels. The synthetic dataset is split into groups according to texture and lighting conditions. The synthetic dataset collection setting is available in [18]. To obtain the ground truth bounding boxes of the lumen, we used the ground truth depth

data provided the synthetic dataset. The depth map is clipped at 3/4 depth from the nearest depth value to segment the lumen. The result is then converted to rectangular bounding boxes. The second set of videos used in our study was acquired with a plastic phantom, an off-the-shelf full HD camera (MISUMI SYT, 1920×1080, 30Hz, field of view of 140 degrees) and a colon model used for training medical professionals. The model is made from plastic and mimicks 1-to-1 the anatomy of the human colon, including internal diameter, and overall length haustral folds (small, segmented pouches of the bowel) to yield accurately simulated images from an optical colonoscope. Creating the haustral folds with this model does not require inflation with air. The camera is connected to a shaft used to navigate inside the plastic colon-rectum tube forward and backward. The external diameter of the camera is 7mm, including light illumination and lens. The number of frames generated from plastic phantom video is 2,042. The annotations of labelled frames in this dataset have been conducted manually using LabelImg[3] software by drawing the bounding boxes around the lumen.

DL experimental settings. We used Faster R-CNN [19] as a fully supervised baseline algorithm in our experiments. Our model and baseline model were trained for 32,000 iterations on plastic phantom data, and for 52,000 iterations on synthetic data as the size of video data varies. The size of the batch was set to 8 with stochastic gradient descent SGD with an initial learning rate of 10^{-2} with momentum 0.9 and weight decay 10^{-3}, which decays by dividing by 10 at iterations 36,000 and 48,000 on synthetic data and 18000 and 28,000 iterations on plastic phantom frames. We also set the unsupervised weight to $\alpha = 2$. The confidence threshold score is set to 0.8 in the inference phase. The models are implemented using Pytorch and trained on an Nvidia RTX A6000 GPU with a memory of 48GB. The implementation of Faster R-CNN with Res50 and hyper-parameter setting are based on the MMDetection library [3]. For data augmentation, we follows the same augmentation schemes applied in FixMatch [22] including colour transformations, translation with translation ratio of $(0, 0.1)$, rotation with angle $(0, 30)$, shifting with angle $(0, 30)$, cut-out [7] with ratio $(0.05, 0.2)$ and number of regions [1, 5].

Results. We evaluated the lumen detection model using both quantitative and qualitative measurements. In terms of qualitative evaluation, we summarise in Fig. 2 a comparison of the semi-supervised model against the baseline model on both video types. For quantitative analysis, shown in Table 1, we use the typical object detection metrics, including average precision (AP) and average recall (AR) using various Intersection over Union (IoU) threshold scores. Although the semi-supervised model was trained on only 10% of frames, it shows a competitive performance without needing expensive manual annotations. Importantly, our lumen detection runs in 60 ms including post-processing time, meeting interventional time requirements.

Discussion. The non-learning based methods [6, 10, 11, 17] have not reported evaluation performance compared to ground truth in overlapping with the bounding boxes. Recently, authors in [29] used off-the-shelf fully supervised model

[3] https://github.com/tzutalin/labelImg

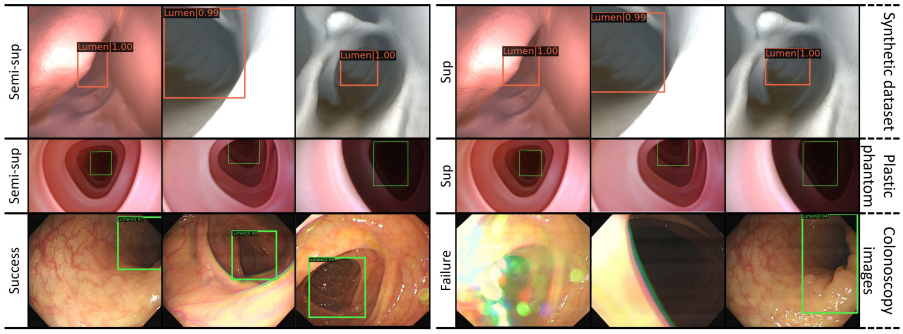

Fig. 2: The qualitative results of lumen detection on test data of synthetic dataset and plastic phantom, respectively. It can be observed that semi-supervised (Semi-sup in figure) prediction accuracy is par with the fully supervised (Sup in figure) model prediction on both datasets. Success and failure cases of the proposed model on real colonoscopy images. Light scattering and low illumination in very challenging conditions are found to affect the prediction.

Yolo3 to localise the lumen targeting to develop semi-automated navigation. They reported an AP of 0.835 with an IoU threshold score of 0.7 achieved by a model trained on 7,147 fully labelled frames captured from plastic phantom. In contrast, the size of our plastic phantom data was only 2,042 frames in total. To further evaluate the generalisation ability, robustness and reliability of the presented model, three colonoscopy videos taken from publicly available dataset [13] that contain a variety of polyps and complex bowel environments are tested on the developed semi-supervised model, pre-trained on the plastic phantom video data. The obtained detection results shown in Fig. 2 (third row) reveal superior performance on unseen real colonoscopy data. Due to the lack of ground truth bounding boxes of these datasets, the lumen detection results on the real colonoscopy videos have been examined by an anonymous survey involving eight senior clinicians. Purpose of this study was to have a qualitative evaluation on the accuracy of the lumen detection. We established a questionnaire showing a rating scale in range (1 - Extremely poor, 5 - Excellent). The average accuracy reported by the clinicians was 4.37 out of 5. These findings demonstrate that the proposed deep feature learning-based approach will be a valuable automated navigation tool to be deployed in a challenging real-time environment during robotics colonoscopy. Furthermore, the integration of automated systems based on large unlabelled data will also significantly reduce the manual data annotations workload and thus reduce costs. Our proposed model has limitations. The real scenario may be more challenging when a colon has an abnormality, such as big polyps, cancer, and diverticula. We target in our future work to systematically study all scenarios, investigate how the model could cope with various conditions and include more ablation studies for experimental settings. More experiments on both two-stage and one-stage detectors will be also conducted to study the generalisation of this method.

Table 1: Comparison of AP and AR for supervised and semi-supervised models on synthetic and phantom data, at different IoU threshold scores. The performance of the Res50 backbone model is also explored here. In addition to the 10% data splitting scheme, the performance of the SSL model is examined in a 5% data splitting scenario.

Data	**Split**	**Backbone**	**IoU**	Semi-supervis		Supervised	
				AP	**AR**	**AP**	**AR**
Synthetic	10%	Res50	0.5-0.95	0.637	0.633	0.621	0.674
			0.5	0.989	0.688	0.978	0.674
			0.75	0.763	0.688	0.713	0.674
		Res101	0.5-0.95	0.668	0.718	0.624	0.691
			0.5	0.989	0.715	0.978	0.681
			0.75	0.807	0.700	0.728	0.689
	5%	Res50	0.5-0.95	0.601	0.662	0.602	0.655
			0.5	0.977	0.667	0.977	0.649
			0.75	0.703	0.669	0.700	0.660
		Res101	0.5-0.95	0.634	0.691	0.616	0.680
			0.5	0.988	0.680	0.976	0.682
			0.75	0.753	0.694	0.700	0.689
Phantom	10%	Res50	0.5-0.95	0.572	0.651	0.562	0.640
			0.5	0.936	0.651	0.950	0.640
			0.75	0.677	0.651	0.638	0.640
		Res101	0.5-0.95	0.567	0.652	0.554	0.632
			0.5	0.933	0.662	0.960	0.637
			0.75	0.628	0.658	0.613	0.632
	5%	Res50	0.5-0.95	0.518	0.600	0.497	0.603
			0.5	0.948	0.604	0.925	0.609
			0.75	0.521	0.600	0.478	0.606
		Res101	0.5-0.95	0.470	0.570	0.452	0.544
			0.5	0.896	0.559	0.950	0.540
			0.75	0.441	0.560	0.325	0.556

4 Conclusions

In this paper, a novel real-time lumen detection and tracking method has been introduced and tested in a plastic phantom, synthetic and real colonoscopy videos. We have introduced the SSL approach toward real-time bound boxes detection of the lumen, allowing for autonomous navigation and thus providing significant benefits in terms of reduced physical burden and demanding the minimum intervention from the operator. Our findings support our key claim that a reliable medical AI-based solution could be established using a small quantity of labelled data combined with other unlabelled data. A paradigm shift like this might pave the way for intelligent robot-assisted diagnosis and treatment.

Acknowledgements This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant number EP/W00433X/1.

References

1. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21660, accessed: 2022-02-27
2. Ahmad, O.F., Soares, A.S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., Stoyanov, D., Chand, M., Lovat, L.B.: Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. The lancet Gastroenterology & Hepatology **4**(1), 71–80 (2019). https://doi.org/10.1016/S2468-1253(18)30282-6
3. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
4. Chen, P.J., Lin, M.C., Lai, M.J., Lin, J.C., Lu, H.H.S., Tseng, V.S.: Accurate classification of diminutive colorectal polyps using computer-aided analysis. Gastroenterology **154**(3), 568–575 (2018). https://doi.org/10.1053/j.gastro.2017.10.010
5. Citarda, F., Tomaselli, G., Capocaccia, R., Barcherini, S., Crespi, M., Group, I.M.S., et al.: Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. Gut **48**(6), 812–815 (2001). https://doi.org/10.1136/gut.48.6.812
6. Ciuti, G., Visentini-Scarzanella, M., Dore, A., Menciassi, A., Dario, P., Yang, G.Z.: Intra-operative monocular 3d reconstruction for image-guided navigation in active locomotion capsule endoscopy. In: 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob). pp. 768–774. IEEE (2012). https://doi.org/10.1109/BioRob.2012.6290771
7. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
8. Di Natali, C., Beccani, M., Valdastri, P.: Real-time pose detection for magnetic medical devices. IEEE Transactions on Magnetics **49**(7), 3524–3527 (2013). https://doi.org/10.1109/TMAG.2013.2240899
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 580–587 (2014). https://doi.org/10.1109/CVPR.2014.81
10. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.M.M.: Defslam: Tracking and mapping of deforming scenes from monocular sequences. IEEE Transactions on Robotics **37**(1), 291–303 (2021). https://doi.org/10.1109/TRO.2020.3020739
11. Liu, J., Subramanian, K.R., Yoo, T.S.: An optical flow approach to tracking colonoscopy video. Computerized Medical Imaging and Graphics **37**(3), 207–223 (2013). https://doi.org/10.1016/j.compmedimag.2013.01.010
12. Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A.: Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE Transactions on Medical Imaging **39**(11), 3429–3440 (2020)
13. Ma, Y., Chen, X., Cheng, K., Li, Y., Sun, B.: Ldpolypvideo benchmark: A large-scale colonoscopy video dataset of diverse polyps. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 387–396. Springer (2021)
14. Manfredi, L.: Endorobots for colonoscopy: Design challenges and available technologies. Frontiers in Robotics and AI p. 209 (2021). https://doi.org/10.3389/frobt.2021.705454

15. Manfredi, L., Capoccia, E., Ciuti, G., Cuschieri, A.: A soft pneumatic inchworm double balloon (spid) for colonoscopy. Scientific Reports **9**(1), 1–9 (2019). https://doi.org/10.1038/s41598-019-47320-3
16. Miguel, M.N., Inaqui, F.U., Cristina, C., Gérard, G., Michel, D., Marie-Georges, L., Thierry, P., Horst, N., Michael, P., Guido, C., et al.: Capsule endoscopy versus colonoscopy for the detection of polyps and cancers. Cancéro Digest (2009)
17. Prendergast, J.M., Formosa, G.A., Heckman, C.R., Rentschler, M.E.: Autonomous localization, navigation and haustral fold detection for robotic endoscopy. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 783–790. IEEE (2018). https://doi.org/10.1109/IROS.2018.8594106
18. Rau, A., Edwards, P., Ahmad, O.F., Riordan, P., Janatka, M., Lovat, L.B., Stoyanov, D.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. International Journal of Computer Assisted Radiology and Surgery **14**(7), 1167–1176 (2019). https://doi.org/10.1007/s11548-019-01962-w
19. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems **28** (2015)
20. Ruijsink, B., Puyol-Antón, E., Li, Y., Bai, W., Kerfoot, E., Razavi, R., King, A.P.: Quality-aware semi-supervised learning for cmr segmentation. In: International Workshop on Statistical Atlases and Computational Models of the Heart. pp. 97–107. Springer (2020)
21. Scaradozzi, D., Zingaretti, S., Ferrari, A.: Simultaneous localization and mapping (slam) robotics techniques: a possible application in surgery. Shanghai Chest **2**(1) (2018). https://doi.org/10.21037/shc.2018.01.01
22. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in Neural Information Processing Systems **33**, 596–608 (2020)
23. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
24. van der Stap, N., van der Heijden, F., Broeders, I.A.: Towards automated visual flexible endoscope navigation. Surgical Endoscopy **27**(10), 3539–3547 (2013). https://doi.org/10.1007/s00464-013-3003-7
25. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems **30** (2017)
26. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10687–10698 (2020)
27. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3060–3069 (2021)
28. Yalniz, I.Z., Jégou, H., Chen, K., Paluri, M., Mahajan, D.: Billion-scale semi-supervised learning for image classification. arXiv preprint arXiv:1905.00546 (2019)
29. Yen, S.Y., Huang, H.E., Lien, G.S., Liu, C.W., Chu, C.F., Huang, W.M., Suk, F.M.: Automatic lumen detection and magnetic alignment control for magnetic-assisted capsule colonoscope system optimization. Scientific Reports **11**(1), 1–10 (2021). https://doi.org/10.1038/s41598-021-86101-9