A Self-supervised Approach for Detecting the Edges of Haustral Folds in Colonoscopy Video

Wenyue Jin[1(✉)], Rema Daher[2], Danail Stoyanov[2], and Francisco Vasconcelos[2]

[1] University College London, London, England
ucabwj0@ucl.ac.uk
[2] WEISS Centre, University College London, London, England

Abstract. Providing 3D navigation in colonoscopy can help decrease diagnostic miss rates in cancer screening by building a coverage map of the colon as the endoscope navigates the anatomy. However, this task is made challenging by the lack of discriminative localisation landmarks throughout the colon. While standard navigation techniques rely on sparse point landmarks or dense pixel registration, we propose edges as a more natural visual landmark to characterise the haustral folds of the colon anatomy. We propose a self-supervised methodology to train an edge detection method for colonoscopy imaging, demonstrating that it can effectively detect anatomy related edges while ignoring light reflection artifacts abundant in colonoscopy. We also propose a metric to evaluate the temporal consistency of estimated edges in the absence of real groundtruth. We demonstrate our results on video sequences from the public dataset HyperKvazir. Our code and pseudo-groundtruth edge labels are available at https://github.com/jwyhhh123/HaustralFold_Edge_Detector.

Keywords: Colonoscopy · Scene understanding · Edge detection · Landmark detection

1 Introduction

Reconstructing 3D gastrointestinal (GI) tract maps from endoscopy videos is a research challenge receiving increasing attention in recent years [4]. In the context of colon cancer screening, real-time 3D reconstruction would enable monitoring which surfaces have already been inspected [13,14], making it easier to ensure complete coverage and reduce the chance of missing polyps [18]. It would also enable complete reporting, associating polyps with precise colon map locations.

Simultaneous Localization and Mapping (vSLAM) is a popular algorithm framework that has been translated to colonoscopy 3D reconstruction [6,17].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-44992-5_6.

B. Bhattarai et al. (Eds.): DEMI 2023, LNCS 14314, pp. 56–66, 2023.
https://doi.org/10.1007/978-3-031-44992-5_6

However, we are still far from reliably reconstructing entire colons in real cases due to multiple imaging challenges. The majority of established methods builds 3D maps relying on the detection of point landmarks in the visualised scene across different frames. In colonoscopy, however, the detection and matching of point landmarks are extremely challenging due to scene textures being very similar, fast camera motions, abundant presence of light reflections, blur, and multiple types of occlusions.

While there is some research towards making point landmark detection more reliable in endoscopy [3], other alternatives involve bypassing the detection of points altogether. Some works perform registration of different frames by directly estimating depth [17] or relative motion [20] using end-to-end deep learning networks. The main challenge here is obtaining the necessary training data. Using virtual simulation has been suggested to train such algorithms [21], however, there is still a gap in generalising its results to real images.

A different alternative to bypass point landmark detection would be to focus on detecting scene edges instead. The colon anatomy has clearly visible and identifiable edges corresponding to its haustral folds (Fig. 1). While edge detection has seen significant progress in computer vision [19], there has been very little investigation on its application to endoscopy. Therefore, we introduce the following contributions:

- We introduce a method to detect haustral fold edges in colonoscopy based on the DexiNed architecture [19]. To the best of our knowledge, it's the first time this problem has been investigated.
- Given the inexistence of groundtruth for colonoscopy edge detection, we propose a combination of transfer learning and self-supervision to train our method.
- We propose an unsupervised evaluation process to measure the temporal consistency of edge predictions in continuous video frames.
- We will release both our code and pseudo-groundtruth edge masks for a subset of the public dataset HyperKvazir.

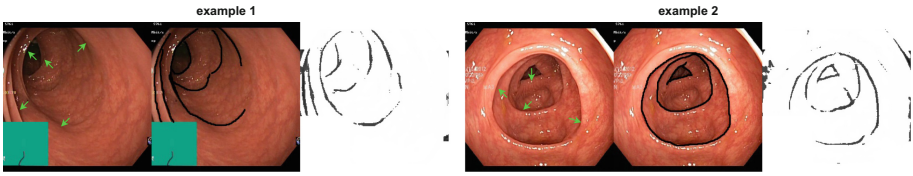

Fig. 1. We aim at detecting haustral folds (denoted by green arrows) in colonoscopy video. Formulating the problem as edge detection, these are circular contours on the colon wall (denoted by black lines). Sample predictions from our method are provided in black and white masks. (Color figure online)

2 Related Work

While most of the classic 3D navigation and reconstruction literature deals with rigid scenes containing unique and easy to recognise visual landmarks, in endoscopy there are two striking differences. The first one is the presence of deformable tissue. A few works have extended visual SLAM to explicitly model deformation of the 3D scene over time [12,22]. The second difference is that it is much more challenging to detect and track reliable landmarks on the GI tract due to simple tissue textures, frequent camera blur, light reflections and other dynamic occlusions. This paper will focus on this latter challenge which we now review in more detail.

Endoscopic scenes contain wet tissues illuminated by a close-range, moving light source. This produces abundant specular light reflections on tissue surfaces and makes it difficult to find landmarks with stable visual appearance. One approach to tackle this is to detect and inpaint specular reflections prior to landmark detection and matching [7]. A large amount of literature is dedicated to detection, filtering and inpainting of specular reflections in surgery [1,9,16].

The different visual appearance of endoscopic scenes presents a very specific domain shift in comparison to well established applications (e.g. outdoors/indoors human-made environments), and therefore machine learning approaches have been useful in bridging this gap. The SuperPoint [8] feature detector can be fine-tuned on endoscopy scenes in a self-supervised way [3], optimising its performance to this particular environment. There are a few other recent deep learning point feature detector alternatives that to the best of our knowledge have not been tried on endoscopy scenes [23,25,26].

Notably, there has been little investigation into the detection of features with other shapes than points. In the context of colonoscopy, this would be a promising direction since the colon is characterised by haustral folds, i. e. thin, ring-shaped structures on its surface (Fig. 1). A recent work has investigated the semantic segmentation of haustral folds [15]. However, we show that its results are still limited and inconsistent when applied to sequences of consecutive frames. We believe there is intrinsic ambiguity in labelling segmentations of these folds, as they do not have a well defined contour in the regions where they join the colon wall. Therefore, we propose to focus instead exclusively on the well defined portion of haustral fold contours using edge detection.

There have been recent advances in performing edge detection with deep learning architectures [19,24]. While pre-trained models are publicly available, these have been trained for general purpose vision, and we show in this paper that they are extremely sensitive to specular reflections. Furthermore, these methods have been trained in a fully supervised fashion, requiring either manually edge labels or proxy edges from semantic segmentation labels. While it would be a burdensome task to produce colonoscopy edge labels in sufficient numbers, we focus instead on self-supervised transfer learning.

3 Methodology

We aim at performing classification of each pixel in colonoscopic images as either edge or not-edge. Our target edges result from the colon shape (i.e. contours of the haustral folds) and not from its surface texture (i.e. vessels, shadows, reflections, etc).

Method Outline. As a baseline we start from the DexiNed model [19] which is a state-of-the-art edge detector trained on non-medical image data. This network is a sequence of 6 convolutional blocks, each of them performing pixel-dense edge detection at different image scales. Using skip connections and up-sampling, these 6 detections are fused into a final multi-scale edge detection result. The network is originally trained with a modified BDCN loss [10] in a fully supervised manner, using manually drawn edges as groundtruth labels.

A pre-trained model of DexiNed is publicly available, and we verify that it is able to detect haustral folds in colonoscopy videos. However, it also produces a significant amount of other false positive detections, mostly artifacts from illumination patterns. While DexiNed is pre-trained in a fully supervised manner, we aim at improving its results on endoscopy data without any additional groundtruth labels available.

Our first observation is that false positive detections can be removed by pre-processing the videos with a temporal specularity inpainting method [7]. In [7], a spatial-temporal transformer is used as a generator within a GAN structure to inpaint specular occlusions. While this produces very appealing results, unfortunately the pre-processing step restricts its usage to offline inference. This is because reliable inpainting results require processing a window of both past and future frames in a single inference step to take advantage of temporal cues. Furthermore such a pipeline would require running two different networks at inference time which is computationally sub-optimal.

To obtain a single model capable of online operation in an end-to-end fashion, we will leverage edges generated with offline pre-processing as pseudo-groundtruth labels to fine-tune DexiNed in a self-supervised manner.

Training Pipeline. Our training methodology is summarised in Fig. 2. We initialise the network with the weights from the original pre-trained DexiNed model, and then fine-tune it on endoscopy video. Our training procedure differs from [19] in the following aspects: (1) Instead of manually annotated groundtruth, we automatically generate pseudo-groundtruth labels with offline processing. (2) Instead of BDCN, we use a mean squared error (MSE) loss, as we empirically verified better results. (3) We train the network on batches of consecutive video frames rather than independent photos. (4) We also add a triplet loss term to improve temporal consistency in continuous video inference.

For a given set of training video clips $c = 1, ..., C$, we generate a set of pseudo-groundtruth label masks $G_{c,t}$ for all frames $X_{c,1}, ..., X_{c,T_c}$ in three steps. First, we pre-process all frames with the inpainting method from [7]. Secondly, we run

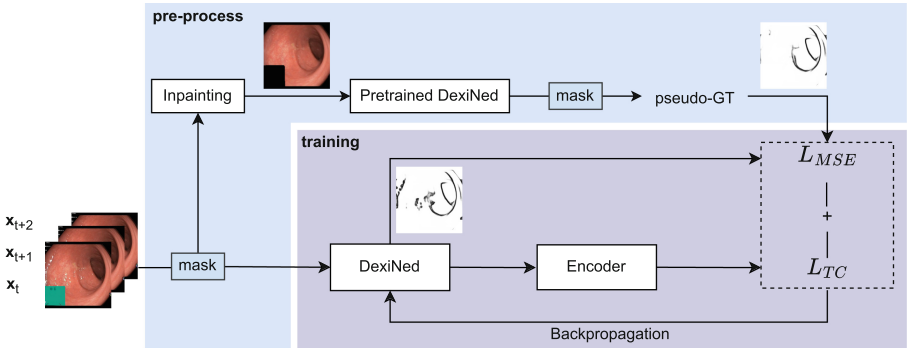

Fig. 2. Model fine-tuning. Pseudo-groundtruth labels are predicted by the pretrained DexiNed on the inpainted images in the pre-processing step. DexiNed is then further trained with a loss combining the pixel-wise loss L_{MSE} and consistency loss L_{TC}. The encoder uses a SegNet model to produce embedding vectors.

the pre-trained DexiNed model on the inpainted training data, generating clean edges of haustral folds. As a last step, we apply a mask to remove any edges resulting from field-of-view and interface overlays typically present in endoscopy images, resulting in the pseudo-groundtruth masks $G_{c,t}$.

Loss Function. We train our model with a loss combining two terms weighted by a parameter γ

$$L = \gamma L_{MSE} + (1 - \gamma) L_{TC} \tag{1}$$

L_{MSE} is the mean squared error between edge predictions $E_{c,t}$ and respective pseudo-labels $G_{c,t}$

$$L_{MSE} = \frac{1}{P} \sum_{c=1}^{C} \sum_{t=1}^{T_c} \sum_{i=1}^{I} \sum_{j=1}^{J} \left(E_{c,t}(i,j) - G_{c,t}(i,j) \right)^2 \tag{2}$$

where I, J are respectively the vertical and horizontal image resolution and P is the total number of pixels in the training data.

L_{TC} is a triplet loss that measures temporal consistency. We take edge predictions from 3 consecutive frames ($E_{c,t}, E_{c,t+1}$, $E_{c,t+2}$) and obtain their lower dimensional embedding vectors with an encoder $\psi()$. We use the encoder from SegNet [2], pre-trained on the Cars dataset[1] The triplet loss is then calculated:

$$\begin{aligned} L_{TC} = \sum_{c=1}^{C} \sum_{t=1}^{T_c - 2} max(\|\psi(E_{c,t}) - \psi(E_{c,t+1})\|_2 - \\ \|\psi(E_{c,t}) - \psi(E_{c,t+2})\|_2 + \beta, 0) \end{aligned} \tag{3}$$

where β is a pre-defined margin parameter. In triplet loss terminology, $E_{c,t}$ represents the anchor, $E_{c,t+1}$ the positive sample, and $E_{c,t+2}$ the negative sample.

[1] The pretrained SegNet is available on https://github.com/foamliu/Autoencoder.

Evaluation. Edge predictions are typically evaluated by comparison against groundtruth labels, through Optimal Dataset Scale (ODS), Optimal Image Scale (OIS) and Average Precision (AP) [19,24]. In addition, our main motivation is to investigate edges as alternative features for navigation, and therefore we aim at temporally consistent estimations that can be further registered in video sequences for camera motion estimation. To this end, we also propose an unsupervised temporal consistency metric based on [27] that was originally introduced for semantic segmentation.

Our evaluation method is summarised in Fig. 3. We measure the temporal consistency TC_t^{t+1} of two independent edge predictions E_t, E_{t+1} by first warping E_t into E'_t using optical flow then measuring the overlap between E'_t and E_{t+1}.

We assume that edge predictions E'_t and E_{t+1} are binarised with a threshold T_1. For optical flow we use FlowNet 2.0 [11]. While [27] computes intersection over union (IoU) between E'_t and E_{t+1} we find this is not adequate for dealing with thin edges. Small spatial shifts in edge predictions result in drastic IoU decrease without it necessarily corresponding to a drastic decrease in edge consistency. Instead, we apply a distance transform to both $E'_{c,t}$ and $E_{c,t+1}$, generating grayscale fields with intensity values representing the distance to the closest edge. The distance fields are binarised with a threshold T_2, resulting in masks D'_t,D_{t+1} denoting all pixels with a distance smaller or equal to T_2 from edges in E'_t,E_{t+1} respectively. The temporal consistency TC_t^{t+1} is a class-weighted IoU between D'_t and D_{t+1}. We weight classes based on their frequency in the image, due to the extreme imbalance between edge and not-edge pixels. Finally, the metric is averaged on all pairs of consecutive frames in the test data.

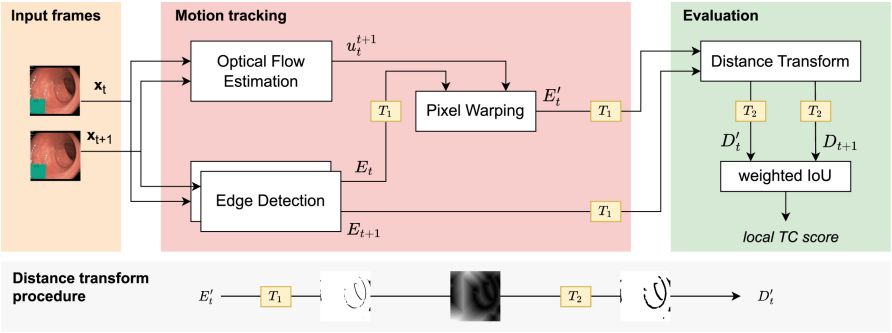

Fig. 3. The framework of consistency evaluation. The motion tracking block produces a pair of edge-maps E'_t and E_{t+1} aligned via optical flow (FlowNet 2.0). The overlap of aligned edge-maps is measured as the class-weighted IoU of binarised distance fields D'_t and D_{t+1}. These distance fields represent pixels within a distance T_2 to the edge predictions.

4 Experiments

Experimental Setup. We train and test our model on a subset of the Hyper-Kvasir dataset [5], defined as all 31 videos of lower GI with adequate bowel preparation (i.e. labelled as BBPS 2-3). We split the data into training, validation, and test with respectively 12, 8, and 11 videos. The images contain black margins and often an endoscope pose display on the lower left corner that produce irrelevant edge detections. We mask out these regions for all images. To compute temporal consistency metrics, we use the totality of the test video data. For comparison against groundtruth, we manually annotated a sparse sub-set of 78 randomly selected images from the test data.

Our method is implemented in Pytorch 1.12.1 with an Intel i7 CPU with 3 GHz and an Nvidia 3090 GPU. Video frames are cropped and resized to 256×256. DexiNed is trained with a RMSprop optimiser with $\alpha = 0.99$ and $\epsilon = 1 \times 10^{-8}$, using a constant learning rate $\eta = 1 \times 10^{-8}$. We use a triplet loss margin $\beta = 1$. A threshold $T_1 = 240$ is set to binarise edge-maps. We use $T_2 = 5$ for model evaluation. We used two-stage training where all models are trained with MSE loss for 5 epochs, followed by 5 epochs of our complete loss in Eq. 1.

Experimental Results Figure 4 displays qualitative results for our model, pseudo-groundtruth, and baselines for a sample sequence of 4 frames. Our model is able to significantly reduce the number of false positive detections caused by highlight reflections. This is a combined effect of the pseudo-groundtruth with temporal consistency (i.e. reflections are less consistent than haustral folds). We note that our method is able to capture the outer edge (see red box) which was not visible either in pre-trained DexiNed or pseudo-groundtruth. We also display results of Foldit [15] for the same sequence, which produces temporally inconsistent fold segmentations that also generally provide less detail about the scene.

In Table 1 we report the temporal consistency (TC), the average percentage of detected edge pixels for each of the tested methods, and also conventional edge accuracy metrics [19,24] ODS, OIS and AP. Our method has higher TC score than all others, including pseudo-groundtruth. This can be explained by the effect of the triplet loss. On average our method detects fewer edge pixels than others which in part is explained by the reduced number of false positive reflection detections (when compared to pre-trained DexiNed) and also due to its thinner edge predictions (when compared to pseudo-groundtruth). In terms of groundtruth evaluation, we observe a comparable performance to the pretrained model when evaluating edge detection metrics. The significantly higher scores obtained for the pseudo-groundtruth validate the reliability of our pseudo-labels. We also highlight that the FoldIt quantitative results should be interpreted with caution (we present them for the sake of completeness) as the detected regions are much larger than our proposed edges. However, its lower TC is consistent with the clearly visible temporal inconsistencies in Fig. 4. In Table 2, we show an ablation of the loss function weight γ. $\gamma = 1.0$ corresponds to using the MSE loss alone, which significantly reduces the TC score.

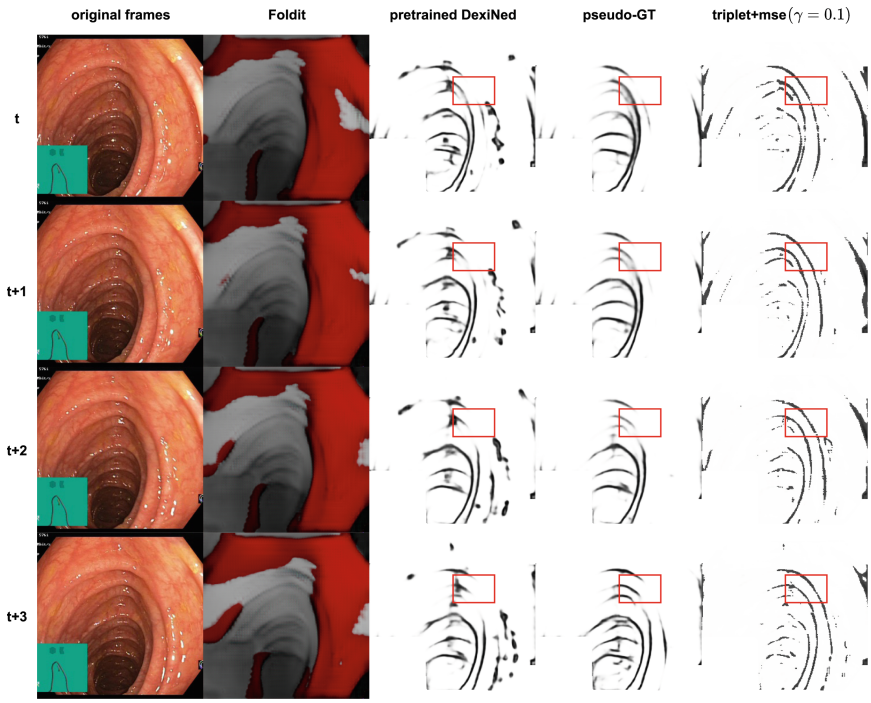

Fig. 4. Edge detection results on four consecutive frames. The predictions in the last column are made by our method. We highlight the red box where significant differences between methods and pseudo-groundtruth can be visualised. (Color figure online)

Table 1. Temporal Consistency (TC), edge pixel rate and results of edge detection metrics (ODS, OIS and AP). We note that Foldit is an image segmentation model (rather than edge detection), which explains the significant differences.

ID	Method	*TC* mean	*TC* std	edge pixel rate	ODS	OIS	AP
1	pretrained DexiNed	0.8840	0.0244	0.1564	0.6332	0.6613	0.5258
2	pseudo-GT	0.9028	0.0172	0.1300	**0.7271**	**0.7556**	**0.6704**
3	Ours (γ=.1)	**0.9348**	**0.0107**	0.0350	0.6491	0.6668	0.5145
4	Foldit	0.8708	0.0359	**0.4976**			

We must note that, as with any unsupervised metric, TC values cannot be analysed in a vacuum. In extreme, a method that never predicts any edge has the highest TC score but this is undesirable. Therefore we should also make sure edge pixel rates are not approaching zero. Our method has an edge pixel rate of 3.5% which is still deemed reasonable for the given data. We note that it is significantly lower than other methods due to detecting thinner edges.

Table 2. Ablation of loss weight γ. All values have similar TC except for $\gamma = 1$ (MSE).

ID	Method	*TC* mean	*TC* std	edge pixel rate
1	triplet+mse (γ=.1)	**0.9348**	**0.0107**	0.0350
2	triplet+mse (γ=.3)	0.9314	0.0111	0.0357
3	triplet+mse (γ=.5)	0.9278	0.0112	0.0351
4	triplet+mse (γ=.7)	0.9310	0.0110	0.0336
5	triplet+mse (γ=.9)	0.9268	0.0117	0.0409
6	triplet+mse (γ=1.0)	0.8445	0.0259	**0.4080**

5 Conclusions

We demonstrate that end-to-end detection of haustral fold edges in colonoscopy videos is feasible and can be made robust to the abundant reflection artifacts present in these scenes with a simple self-supervised training pipeline. We believe these are stable and consistent features across multiple views that can be exploited for colonoscopy video navigation and place recognition, but so far have been underexplored. While our method shows promising qualitative results and temporal consistency, future work should evaluate these features in downstream tasks such as endoscope motion estimation, 3D reconstruction, and place recognition.

Acknowledgments. This work was supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) at UCL (203145Z/16/Z) and the H2020 FET EndoMapper project (GA863146). This work was partially carried out during the MSc in Robotics and Computation graduate degree at the Computer Science Department, UCL.

References

1. Ali, S., et al.: A deep learning framework for quality assessment and restoration in video endoscopy. Med. Image Anal. **68**, 101900 (2021)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. CoRR abs/ arXiv: 1511.00561 (2015)
3. Barbed, O.L., Chadebecq, F., Morlana, J., Martínez-Montiel, J., Murillo, A.C.: Superpoint features in endoscopy. arXiv preprint arXiv:2203.04302 (2022)
4. Batlle, V.M., Montiel, J.M., Tardós, J.D.: Photometric single-view dense 3d reconstruction in endoscopy. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4904–4910. IEEE (2022)
5. Borgli, H., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci. Data **7**(1), 283 (2020). https://doi.org/10.1038/s41597-020-00622-y

6. Chen, R.J., Bobrow, T.L., Athey, T., Mahmood, F., Durr, N.J.: Slam endoscopy enhanced by adversarial depth prediction. arXiv preprint arXiv:1907.00283 (2019)
7. Daher, R., Vasconcelos, F., Stoyanov, D.: A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence (2022). https://doi.org/10.48550/ARXIV.2203.17013
8. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: self-supervised interest point detection and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 224–236 (2018)
9. Garcia-Vega, A., et al.: Multi-scale structural-aware exposure correction for endoscopic imaging. arXiv preprint arXiv:2210.15033 (2022)
10. He, J., Zhang, S., Yang, M., Shan, Y., Huang, T.: Bi-directional cascade network for perceptual edge detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3828–3837 (2019)
11. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. CoRR abs/ arXiv: 1612.01925 (2016)
12. Lamarca, J., Parashar, S., Bartoli, A., Montiel, J.: Defslam: tracking and mapping of deforming scenes from monocular sequences. IEEE Trans. Rob. **37**(1), 291–303 (2020)
13. Ma, R., et al.: Real-Time 3D reconstruction of colonoscopic surfaces for determining missing regions. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11768, pp. 573–582. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32254-0_64
14. Ma, R., et al.: Rnnslam: reconstructing the 3d colon to visualize missing regions during a colonoscopy. Med. Image Anal. **72**, 102100 (2021)
15. Mathew, S., Nadeem, S., Kaufman, A.: FoldIt: haustral folds detection and segmentation in colonoscopy videos. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 221–230. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_21
16. Monkam, P., Wu, J., Lu, W., Shan, W., Chen, H., Zhai, Y.: Easyspec: automatic specular reflection detection and suppression from endoscopic images. IEEE Trans. Comput. Imaging **7**, 1031–1043 (2021)
17. Ozyoruk, K.B., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med. Image Anal. **71**, 102058 (2021)
18. Pickhardt, P.J., Taylor, A.J., Gopal, D.V.: Surface visualization at 3d endoluminal ct colonography: degree of coverage and implications for polyp detection. Gastroenterology **130**(6), 1582–1587 (2006)
19. Poma, X.S., Sappa, Á.D., Humanante, P., Akbarinia, A.: Dense extreme inception network for edge detection. CoRR abs/ arXiv: 2112.02250 (2021)
20. Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D.: Bimodal camera pose prediction for endoscopy. arXiv preprint arXiv:2204.04968 (2022)
21. Rau, A., et al.: Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy. Int. J. Comput. Assist. Radiol. Surg. **14**, 1167–1176 (2019)
22. Rodríguez, J.J.G., Montiel, J.M., Tardós, J.D.: Tracking monocular camera pose and deformation for slam inside the human body. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5278–5285. IEEE (2022)
23. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4938–4947 (2020)

24. Soria, X., Riba, E., Sappa, A.: Dense extreme inception network: towards a robust cnn model for edge detection. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1912–1921. IEEE Computer Society, Los Alamitos, CA, USA (Mar 2020). https://doi.org/10.1109/WACV45572.2020.9093290, https://doi.ieeecomputersociety.org/10.1109/WACV45572.2020.9093290
25. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: detector-free local feature matching with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8922–8931 (2021)
26. Tang, J., Ericson, L., Folkesson, J., Jensfelt, P.: Gcnv2: efficient correspondence prediction for real-time slam. IEEE Robotics Autom. Lett. **4**(4), 3505–3512 (2019)
27. Varghese, S., et al.: Unsupervised temporal consistency metric for video segmentation in highly-automated driving. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1369–1378 (2020). https://doi.org/10.1109/CVPRW50498.2020.00176