Endoscopic polyp segmentation using a hybrid 2D/3D CNN

Anonymous

Anonymous

Abstract. Colonoscopy is the gold standard for early diagnosis and pre-emptive treatment of colorectal cancer by detecting and removing colonic polyps. Deep learning approaches to polyp detection have shown potential for enhancing polyp detection rates. However, the majority of these systems are developed and evaluated on static images from colonoscopies, whilst applied treatment is performed on a real-time video feed. Non-curated video data includes a high proportion of low-quality frames in comparison to selected images but also embeds temporal information that can be used for more stable predictions. To exploit this, a hybrid 2D/3D convolutional neural network architecture is presented. The network is used to improve polyp detection by encompassing spatial and temporal correlation of the predictions while preserving real-time detections. Extensive experiments show that the hybrid method outperforms a 2D baseline. The proposed architecture is validated on videos from 46 patients. The results show that real-world clinical implementations of automated polyp detection can benefit from the hybrid algorithm.

Keywords: Colonoscopy · Polyp Detection · Computer Aided Diagnosis.

1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide accounting for 10% of all forms of cancer [4] but early diagnosis and treatment can significantly improve the associated prognosis. Colonoscopy is the gold standard colon screening procedure for early detection, during which the bowel is visually inspected for polyps using an endoscope [12]. Unfortunately colonoscopy is highly operator dependent, with high reported polyp miss rates and associated interval cancers [14].

Computer-aided polyp detection (CAD) systems aiming to assist endoscopists with automatic polyp identification from video have been researched for several decades but significant clinical progress has been reported only in recent years [6, 15, 16]. In particular, approaches based on Convolutional Neural Networks (CNNs) [3, 13, 16] have reported robust and promising results [1]. One of the main challenges when developing such detection models is the limited availability of labelled data because full length colonoscopic videos are not usually recorded clinically, whereas still frames are stored in clinical reports enabling still

image databases [2, 8]. While most current CAD systems have been trained and evaluated on still images they are used in endoscopy units where real time videos are used to detect polyps. It is therefore necessary to demonstrate sound performance on videos and address model behavior stability in practical conditions with poor visibility and variability in polyp appearance that might lead to a lack of temporal coherence in consecutive frames yielding short, false predictions [2].

Yet temporal information in endoscopic video can be exploited for more temporally correlated predictions by extracting temporal representations. Recurrent neural networks (RNN), such as long short-term memory (LSTM), 3D CNNs, or two-stream models have demonstrated good results for temporal recognition tasks [5]. In endoscopic CAD, dense 3D networks have been explored, such as C3D to classify endoscopic frames containing polyps [7, 10] and also a 3D Fully Convolutional Network for polyp segmentation [17]. A major challenge remains the problem of training a 3D CNN with a limited number of videos and various strategies have been proposed to overcome this limitation. For example including a module following a CNN's prediction to increase temporal coherence on consecutive frames or a hand-tuned false positive reduction stage appended to the polyp detection model [11]. Tracking algorithms can be combined with detection CNNs to temporally refine results but the re-initialisation of the tracker can be problematic [19]. Recently an approach fusing two CNN streams, one receiving the input frame, and the other one optical flow information, was reported but can suffer from errors in the optical flow estimation [18].

In this paper, we propose a novel hybrid 2D/3D architecture for polyp detection and segmentation in colonoscopic videos. The proposed architecture intrinsically learns spatio-temporal representations from videos. This increases the network's ability to generalize to challenging clinical endoscopic situations with lower quality data or temporally inconsistent data. A 2D neural network is used to extract spatial features and leverages large training databases through transfer learning. The 3D network component ensures temporal consistency in an efficient architecture designed for real time performance. The hybrid method has been quantitatively and qualitative evaluated and bench-marked against a 2D segmentation network. The results show an increase in performance with higher sensitivity, higher specificity, better spatial segmentation and more stable temporal segmentation.

2 Methods

A two-step temporal segmentation algorithm was developed (see Figure 1). The proposed architecture was capable of learning a spatial representation of polyps through the 2D stage, allowing to apply transfer learning from larger 2D datasets. A 3D segmentation stage followed in order to generate temporally coherent polyp segmentation masks.

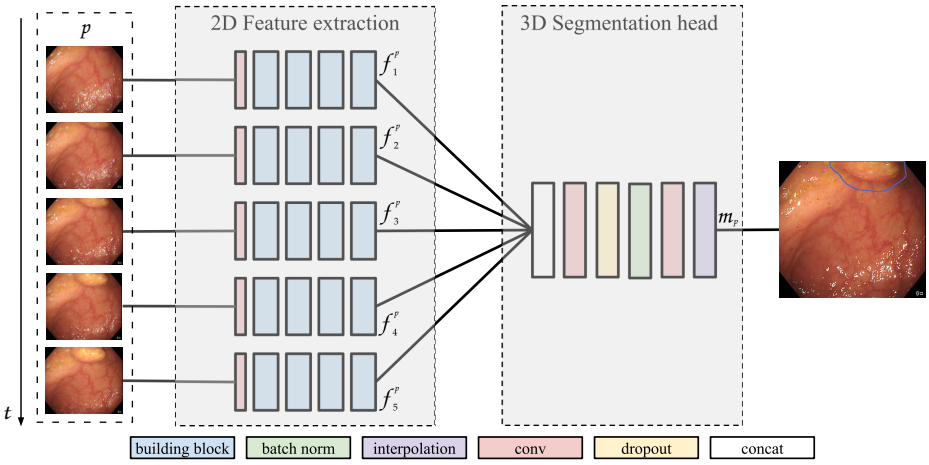

Fig. 1. Architecture of the proposed hybrid segmentation network. A polyp pseudo-batch as input and the corresponding output segmentation are presented.

2.1 Hybrid architecture

The first part of the hybrid model corresponds to the 2D feature extraction. A Resnet-101 architecture was used as a backbone, which included a convolutional layer, followed by four sets of building blocks containing 3, 4, 23 and 3 residual blocks, sequentially. The last fully-connected layer was removed, the output then consisting of a set of 2048 feature maps per image.

The 3D segmentation stage is composed of two 3D convolutional layers, dropout, batch normalisation and an interpolation layer for upsampling (see Figure 1). This structure is an inflated version of the segmentation head from a Fully Convolutional Network (FCN) [9]. The first convolutional layer reduces the number of features by four, applying $[d \times 3 \times 3]$ convolutions, where d is the depth, with a padding and a stride of $[1 \times 1 \times 1]$. The second convolutional layer uses the same stride, no padding, and a kernel of $[d - 2 \times 3 \times 3]$, outputting one channel per class. Note that inherently the temporal depth of the output maps is reduced by four filters in the segmentation head.

An input batch for the network contained N images, composed by P pseudo-batches formed by d consecutive frames each. Each image in a pseudo-batch p was passed through a separate stream of the Resnet backbone extracting a set of spatial features f_i^p. The concatenation step in the segmentation head stacks the features f_i^p for the images in the same pseudo-batch $p \in [1, ..., P]$, where $P = \frac{N}{d}$. The input to the first convolutional layer has a shape of $[P \times 2048 \times d \times w \times h]$, where w and h are the width and height of feature maps. The network generates a probability map m_p per pseudo-batch, corresponding to the image in the middle of the pseudo-batch.

During training, an additional auxiliary segmentation head receives features from the third backbone building block. These feature maps have undergone

fewer pooling steps, and therefore are twice the size of the main backbone's feature maps. Thus, the corresponding segmentation output before the interpolation layer is double the resolution of the main segmentation output. A combination $\mathcal{L}_{aux} + \mathcal{L}_{main}$ of the losses computed from the two segmentation heads is used as the final loss $\mathcal{L}$, allowing to refine the spatial precision of the output.

2.2 Training strategy

A temporal window of $d = 5$ consecutive frames was selected to optimize the balance between new temporal information without sacrificing detection speed. Additionally, $d = 5$ intrinsically yields a single segmentation output from the model as explained in the previous section.

Random sampling of 5000 pseudo-batches was performed on each epoch to minimise overfitting, re-sampling at every new epoch. Data augmentation was applied in such a way as to guarantee identical augmentations within pseudo-batches. The augmentation operations consisted of random affine transformations (rotation, translation and scale) and random colour transformations (brightness, contrast and saturation). Finally, the images were preprocessed by cropping out the video borders, followed by resizing the images to 448 by 448 pixels, and an intensity normalization step.

All available positive images were used during training, and a data mining strategy was adopted for selecting the most beneficial negative images to the training set. An initial model was trained uniquely with positive samples and was used for inference on the available set of negative images (from training procedures), which was shuffled randomly. Images yielding false positives were selected until reaching 15% of the new training set.

Cross-entropy loss was used the experiments and Adagrad for optimisation with a batch size N of 45, and pseudo-batch size P of 9. Two output classes were defined: polyp and no-polyp presence. The epoch with the highest pixel accuracy in the validation set was selected for testing. All models were trained with Pytorch on an NVIDIA Tesla V100 DGXS 32GB GPU. The hybrid network was able to predict at 19 frames per second, ensuring real-time predictions.

3 Experimental results

3.1 Datasets

The data was divided into two separate datasets: *Dataset V* composed of consecutive video frames and, *Dataset S* composed of static images.

Dataset V. A series of 95 videos, from 95 patients, were collected in XX Hospital with an Olympus XX endoscope under ethics REC reference XX/XX/XX. A total of 234 histologically confirmed polyps were extracted into single-polyp video sequences. The frames in these sequences were annotated by expert colonoscopists by drawing bounding boxes around each polyp. Only polyp, white light frames were included. The 25 full-length negative videos were added to the testing set, whereas the 70 procedures containing polyps were randomly split into

training, validation and testing sets on a per-patient basis, avoiding any type of data contamination. 51,426 frames from 173 polyps within 45 procedures were used for training and 2,152 frames from 8 polyps within 4 procedures were used for validation. 20,943 frames from 21 procedures and 53 polyps were used for testing, as well as 542,583 non-polyp frames from the 20 negative procedures.

Dataset S. Static polyp images were gathered from two sources: the publicly available Kvasir dataset [8] composed of 1000 polyp images and corresponding masks, and a dataset containing 833 polyp images collected from reports from XX Hospital under ethics REC reference XX/XX/XX and annotated by expert colonoscopists by drawing bounding boxes around polyps. This set of 1,833 white light, polyp images was solely used for training purposes.

The use of a public test set was not possible as the only test video dataset reported, the ASU-MAYO video dataset, was not available [13].

3.2 Comparison and evaluation metrics

In order to assess the temporal benefits of our model, its comparable 2D network, an FCN with a Resnet101 backbone, was implemented [9]. Whereas the backbone used was identical to the one in the hybrid model, the segmentation head was a deflated version of the hybrid one. In this case, 3D convolutional, batch normalisation and pooling layers were replaced by their 2D corresponding versions, maintaining all parameters. During training, an auxiliary segmentation head was used in the same manner as for the hybrid network. The training strategy and parameters for the baseline model were kept identical to the hybrid model, when possible, to ensure comparison fairness.

Object-wise metrics were used for evaluation, namely sensitivity, precision, and F1-score on videos with polyps, and specificity on non-polyp videos. Further implementation details are available in [2]. Polyp objects were denoted by a rectangle enclosing pixels classified as polyp at a threshold of 0.5. This allowed comparison with the ground truth annotations of rectangular bounding boxes. Dice score was reported on true positive frames to assess the quality of the overlap. Per-polyp sensitivity was also reported, considering a true positive when at least one frame was correctly detected for each polyp.

In order to determine the consistency of the predictions over consecutive frames, temporal coherence (TC) was computed as defined in [2]. Additionally, auto-correlation of masks was measured to assess both temporal and spatial correlation between two consecutive mask predictions. The auto-correlation for a given pixel position over a sequence of masks is defined as:

$$r = \frac{\sum_{i=1}^{N-k}(Y_i - \bar{Y})(Y_{i+1} - \bar{Y})}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \tag{1}$$

where Y_i is the value of a pixel in a certain position, and $\bar{Y}$ is the average of the pixel values in that position over the entire sequence. After obtaining a 2D vector with auto-correlation values per sequence, the average over the x and y axis was computed. The absolute difference with respect to the ground truth auto-correlation was computed, and mean and standard deviation were reported.

3.3 Results and analysis

To establish a baseline, an FCN was trained initialising the backbone weights from ImageNet - referred to as *FCN (ImageNet)*. The training set consisted of images from the *Dataset V* training set and the full *Dataset S*. Furthermore, 10,000 negative images from the training procedures from *Dataset V* were added to the training set using the strategy previously described, by means of an FCN model formerly trained on positive images exclusively. Correspondingly, a hybrid model was trained on the training set from *Dataset V*, and 10,000 negative images, following the negative mining strategy. The hybrid network was trained using the weights from *FCN (ImageNet)* to initialise and freeze the backbone, only training the segmentation head (*Hybrid (FCN)*). Having a common backbone, it was then possible to evaluate the effect of the 3D segmentation head. Table 1 depicts the associated results when tested on the *Dataset V* testing set, where it can be observed that the incorporation of the 3D component caused a general increase in performance, particularly in terms of temporal consistency shown by the high temporal coherence and low distance between the predictions and ground truth auto-correlations.

Table 1. Quantitative evaluation of baseline and proposed methods on Video *Dataset V* (*pp* and *pf* denote per-polyp and per-frame sensitivity, respectively)

Method	Sens (*pp*)(%)	Sens (*pf*)(%)	Spec (%)	Prec (%)	F1 (%)	Dice (%)	Δ A-corr (%)	TC (%)
FCN (ImageNet)	100.00	83.56	83.04	88.11	85.78	69.68	20.50±16.16	79.55
Hybrid (FCN)	100.00	85.66	83.60	93.27	89.30	**74.08**	**11.92±11.97**	84.24
Hybrid (ImageNet)	100.00	**86.14**	**85.32**	**93.45**	**89.65**	73.48	12.24±11.74	**84.64**

The proposed model was also trained initialising the backbone from ImageNet weights and training the full network. This experiment will be referred to as *Hybrid (ImageNet)*. This showed the highest F1-score in Table 1, demonstrating that it is possible to train the hybrid architecture with limited amounts of data. As it can be observed, incorporating temporal components led to a reduction of false positives and negatives in both hybrid models. Figure 2 shows the per frame predictions on one of the non-polyp full colonoscopic withdrawals from the testing set, where it can be seen that the number of false positives is reduced throughout the procedure with the hybrid network compared to the FCN. Although the mapping to a 3D model of the colon is not fully realistic, it gives an indication of the importance of reducing the false positives. In addition to an improvement in sensitivity, the dice score increased considerably on detected polyps with the hybrid model, showing that the quality of the segmentation masks benefited from the temporal component. This can be supported by the increase in auto-correlation, which indicates that consecutive predicted masks presented a higher similarity. Finally, the temporal coherence benefited from the 3D component, suggesting a decrease of short false positive predictions.

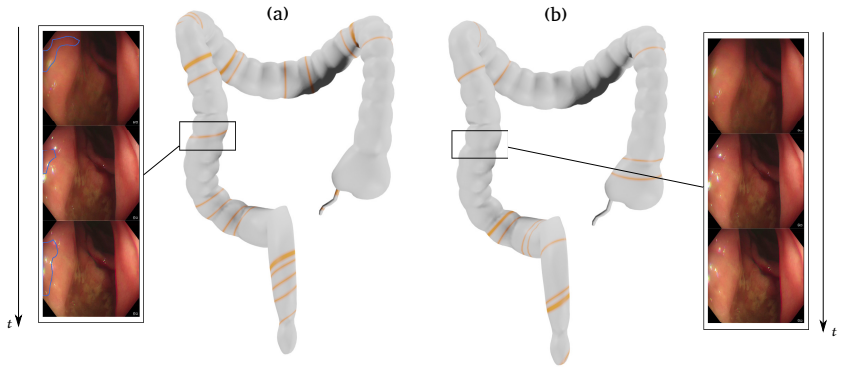

Fig. 2. Prediction timelines for a non-polyp procedure mapped onto a colon model for (a) FCN (ImageNet) and (b) Hybrid (ImageNet), where orange stripes denote false positives. Network outputs are shown as an overlay on a video section.

The enhanced temporal correlation can also be observed on Figure 3. The dice score over a sequence of polyp frames is more stable for the hybrid model, corroborating the increased dice score and auto-correlation. The segmentation examples in Figure 3(bottom) show that both models generated similar outputs on good quality images. However, on blurry frames the FCN yielded false negatives while the hybrid model successfully used information from surrounding frames to predict correctly.

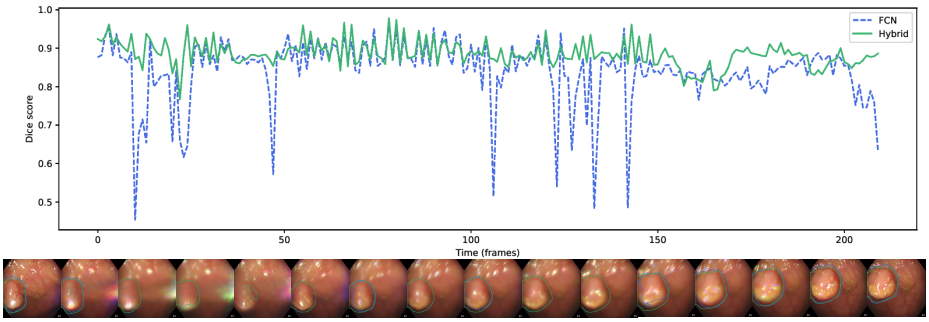

Fig. 3. Results on a polyp sequence showing (top) the dice overlap with the ground truth and (bottom) segmentation outputs for the FCN (blue) and Hybrid (green)

Because the negative images used to train the models reported in Table 1 were not identical due to the negative mining strategy, a data benchmark was performed to show the closest comparison possible between models. Table 2 shows the performance of the models when trained uniquely on positive data. A 2D FCN was trained on *Dataset V* and *Dataset S*, initialising its backbone

from ImageNet weights. A hybrid model was trained by initialising the backbone from this *FCN (ImageNet)*, and training exclusively the segmentation head. A hybrid model was also trained from ImageNet, without freezing any layers. It is important to note that these networks were trained exclusively on positive samples and false positives are to be expected. When compared to the FCN, which was also trained on *Dataset S*, the results presented in Table 2 show that the hybrid model initialised from ImageNet yields poorer results. This can indicate that the model might be overfitting when training only on positive images, a common problem on 3D architectures. However, when initialised from an FCN trained on polyps, the proposed model improved the performance considerably in all aspects. Particularly, a 20.19% rise in specificity was achieved. This shows that the proposed architecture allows to pre-train on still images while benefiting from the temporal stability provided by the 3D segmentation head.

Table 2. Detailed evaluation of network performance when negative data is not included in the training set. Results are reported on the test set from *Dataset V* (*pp* and *pf* denote per-polyp and per-frame sensitivity, respectively).

Method	Sens (*pp*) (%)	Sens (*pf*) (%)	Spec (%)	Prec (%)	F1 (%)	Dice (%)	Δ A-corr (%)	TC (%)
FCN (ImageNet)	100.00	87.77	54.02	87.81	87.80	72.22	17.12±15.18	84.58
Hybrid (FCN)	100.00	**88.88**	**74.18**	**92.73**	**90.76**	**75.06**	**9.77±9.77**	**87.78**
Hybrid (ImageNet)	100.00	85.79	44.54	87.45	86.61	68.27	10.89±11.25	84.42

4 Discussion and Conclusion

A novel hybrid 2D/3D segmentation CNN architecture for polyp detection in colonoscopic videos was developed. As a result of its 2D feature extraction, the hybrid network encompasses the benefits from a 2D architecture, namely spatial representation learning and the potential opportunity to apply transfer learning from a curated dataset of still images. This is particularly beneficial for clinical applications, where large video datasets are challenging to collect and are currently not widely available and hence still image data may be needed to provide strong and diverse representation of the spatial domain. In our method, the 3D segmentation seamlessly incorporates temporal correlation in the results encapsulating learning of spatio-temporal information on smaller video datasets. The overall hybrid architecture is validated on videos from 46 patients, showing an increase in performance, along with higher quality segmentation potential. Future work includes optimization of the length of temporal information and potential handling of video aberration artefacts and expanding testing to larger video datasets.

References

1. 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)
2. Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al.: Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE transactions on medical imaging **36**(6), 1231–1249 (2017)
3. Brandao, P., Zisimopoulos, O., Mazomenos, E., Ciuti, G., Bernal, J., Visentini-Scarzanella, M., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., et al.: Towards a computed-aided diagnosis system in colonoscopy: automatic polyp segmentation using convolution neural networks. Journal of Medical Robotics Research **3**(02), 1840002 (2018)
4. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
6. Hassan, C., Wallace, M.B., Sharma, P., Maselli, R., Craviotto, V., Spadaccini, M., Repici, A.: New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. Gut pp. gutjnl–2019 (2019)
7. Itoh, H., Roth, H.R., Lu, L., Oda, M., Misawa, M., Mori, Y., Kudo, S.e., Mori, K.: Towards automated colonoscopy diagnosis: binary polyp size estimation via unsupervised depth learning. In: International conference on medical image computing and computer-assisted intervention. pp. 611–619. Springer (2018)
8. Jha, D., H. Smedsrud, P., Riegler, M., Halvorsen, P., Johansen, D., de Lange, T., D. Johansen, H.: Kvasir-seg: A segmented polyp dataset. In: Proceedings of the International Conference on Multimedia Modeling (MMM). Springer (2020), https://datasets.simula.no/kvasir-seg/
9. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
10. Misawa, M., Kudo, S.e., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., Ogawa, Y., Maeda, Y., Takeda, K., Ichimasa, K., et al.: Artificial intelligence-assisted polyp detection for colonoscopy: initial experience. Gastroenterology **154**(8), 2027–2029 (2018)
11. Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y.: Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE Journal of Biomedical and Health Informatics (2019)
12. Rex, D.K., Johnson, D.A., Anderson, J.C., Schoenfeld, P.S., Burke, C.A., Inadomi, J.M.: American college of gastroenterology guidelines for colorectal cancer screening 2008. American Journal of Gastroenterology **104**(3), 739–750 (2009)
13. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE transactions on medical imaging **35**(2), 630–644 (2015)

14. Van Rijn, J.C., Reitsma, J.B., Stoker, J., Bossuyt, P.M., Van Deventer, S.J., Dekker, E.: Polyp miss rate determined by tandem colonoscopy: a systematic review. American Journal of Gastroenterology **101**(2), 343–350 (2006)
15. Wang, P., Li, L., Liu, P., Xiao, X., Song, Y., Zhang, D., Li, Y., Xu, G., Tu, M., Xiao, X., et al.: Mo1712 automatic polyp detection during colonoscopy increases adenoma detection: An interim analysis of a prospective randomized control study. Gastrointestinal Endoscopy **87**(6), AB490–AB491 (2018)
16. Wang, P., Xiao, X., Brown, J.R.G., Berzin, T.M., Tu, M., Xiong, F., Hu, X., Liu, P., Song, Y., Zhang, D., et al.: Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nature biomedical engineering **2**(10), 741–748 (2018)
17. Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A.: Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE journal of biomedical and health informatics **21**(1), 65–75 (2016)
18. Zhang, P., Sun, X., Wang, D., Wang, X., Cao, Y., Liu, B.: An efficient spatial-temporal polyp detection framework for colonoscopy video. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI). pp. 1252–1259. IEEE (2019)
19. Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y.: Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern recognition **83**, 209–219 (2018)