An improved automatic system for aiding the detection of colon polyps using deep learning

1st Lishan Cai
Department of Radiology
The Netherlands Cancer Institute
Amsterdam, Netherlands
Oncology and Developmental Biology
Maastricht university
l.cai@nki.nl

2nd Regina Beets-Tan
Department of Radiology
The Netherlands Cancer Institute
Amsterdam, Netherlands
Oncology and Developmental Biology
Maastricht university
r.beetstan@nki.nl

3rd Sean Benson
Department of Radiology
The Netherlands Cancer Institute
Amsterdam, Netherlands
s.benson@nki.nl

***Abstract*—Colorectal cancer is responsible for the most cancer deaths after lung cancer. It has been well-established that early detection and removal of polyps can prevent colorectal cancer. It is therefore essential that automated polyp detection has the highest sensitivity and precision possible in order to detect the most cases and prevent unnecessary treatment. We present a deep learning model based on YOLOv3 that was trained to detect polyps. Training made use of the 39308 images of 78 polyps and 393 completely healthy images from the SUN database. The model was subsequently validated using both the public CVC-clinic and ETIS-Larib datasets containing both standard defintion (SD) and high definition (HD) images. The per-image polyp detection sensitivity(precision) was calculated as 91.5(96.6)% and 86.5(94.2)% for the CVC-clinic and Etis-Larib datasets, respectively. These results represent the best-known performance in the validation datasets in comparison with the results of a recent review.**

***Index Terms*—Colonoscopy, Polyp Detection, Artificial Intelligence, Deep Learning, YOLOv3**

I. Introduction

Colonoscopy is an exam used to detect changes and anomalies in the large intestine (colon) and rectum. It is also regarded as the gold-standard screening test [1]–[3] for colorectal cancer (CRC) and it prevents approximately two-thirds of deaths on the left side of the colon [4]. CRC, which arises from precancerous polyps, is the second leading cause of death in the United States [5]. The National Polyp Study showed that 70%-90% of CRCs are preventable with colonoscopies and complete removal of polyps. Approximately, 85% of interval cancers arise from missed polyps or incompletely removed polyps during colonoscopy [6].

The benefit of colonoscopy for the prevention of CRC relies on the adenoma detection rate. Manual examination by a gastroenterologist currently stands as the first choice for quality measures in screening colonoscopy. However, the detection rates of gastroenterologists vary from 7% to 53%. It is estimated that every 1% increase in detection rate lowers the risk of interval colorectal cancers by 3%–6% [7]. It is necessary to introduce new accurate strategies to increase the polyp detection rate during colonoscopy.

Computer-aided image analysis has the potential to improve polyp detection and attracts widespread attention. In several studies, it shows the promise to reduce the possible missed polyps. It has been reported that one-fourth of neoplastic polyps may be missed on colonoscopy [8] and that more than half of post-colonoscopy CRC may arise from these missed lesions [9]. Moreover, with automatic systems, the polyp detection process is less time-expensive and less resource-consuming. Despite significantly higher polyp detection rate, no improvement in detection of advanced colonic lesions, especially large and significant adenomas or serrated polyps, has been seen with automatic systems and remains a challenge [10].

Recently, Artificial Intelligence (AI) has been reported to speed-up and automate medical image analysis obtaining promising results. Deep learning is the main contributor of the rise in AI in a wide range variety field of application including Computer Vision, Natural Language Processing and medical image analysis [11]. In deep learning approaches, typically a convolution neural network (CNN) is used in order to extract relevant features. The deep learning model often created using transfer learning from a generalised model. In practice, this means that the model has been trained to classify accurately thousands of non-medical images, with a subset of the so-called backobone network refined on the smaller medical image dataset of the use case in question [12].

Advances in transfer learning in years have greatly increased the ability of deep-learning methods to be used in combination with smaller datasets, which is of particular interest for medical imaging where dataset sizes are limited by the number of patients examined subject to the relevant exclusion criteria [13]. Automatic polyp detection has been an active topic for the past years with the utilization of AI, but the performance levels are far from that of the expert gastroenterologist [14]–[16]. The large datasets present for polyp detection provide not only an important means to create models to detect polyps and prevent CRC, but will also provide solid foundations for future transfer learning to use cases in which endoscopy data is not so prevalent, namely cancer regrowth detection and active monitoring [17].

We present a deep-earning algorithm for the automatic detection of polyps during colonoscopy based on transfer

learning of a pre-trained YOLOv3 model [18], which has previously shown promise in the analysis of endoscopic imaging to detect colonic perforation [19] and indeed polyps [20]. We trained our system with one public colonoscopy database and validated the algorithm with two independent datasets. Our results are then compared against the current state-of-the-art [10].

II. Datasets and Methodology

A. Training Dataset

The training dataset used was from Showa University and Nagoya University, referred to as the SUN Colonoscopy Video Database. They used a high-definition endoscope (CF-HQ290ZI and CF-H290ECI; Olympus, Tokyo, Japan), and all colonoscopies were recorded by a high-definition (1008×1158) video recorder (IMH-10; Olympus). Also, all patients were older than 18 years. In total, there were 99 patients with 100 polyps registered. The database contains 49,136 polyp frames [21]. Diagnosis details of the database are summarized in TABLE I.

TABLE I
SUN Colonoscopy Database Details

Pathological Diagnosis	Num	Location	Num
Hyperplastic polyp	7	Right	47
Sessile serrated lesion	4	Left	44
Low grade adenoma	82	Rectum	8
Traditional serrated adenoma	2	-	-
High grade adenoma	4	-	-
Invasive carcinoma	1	-	-

B. Test Dataset B: CVC-clinic Database

The CVC-ClinicDB database includes 612 standard definition still images of 384×288, arising from 29 polyp-positive sequences [22]. In total, there are 646 polyps presented. All the images were acquired from Hospital Clinic of Barcelona, Barcelona, Spain and using an Olympus Q160AL/Q165L colonoscope. The ground truth for each polyp was provided with the format of segmentation masks (see Table II).

C. Test Dataset C: ETIS-Larib Database

The ETIS-Larib is a polyp database that contains 196 high-definition still images with a resolution of 1225×964 of 44 different polyps from 34 sequences [23]. Overall, there are 44 examples of different polyps presented in sizes and viewpoints. Some images have two or three polyps, making the total number of polyp appearances 208. The ground truth was provided in the form of the segmentation mask (see Table II).

TABLE II
Database Summary

Database	Use	Resolution	Patients	Image(polyp)
SUN	train	1008 × 1158	99	49136 HD
CVC-ClinicDB	test	384 × 288	23	612 SD
ETIS-Larib	test	1225 × 966	-	196 HD

D. AI Algorithm

To develop the AI algorithm, we used YOLOv3 without any structural modification. Darknet53 [24] is chosen as the backbone given that it is more performant than Darknet19 but still more efficient than ResNet101 and ResNet152. Darknet53 uses 3×3 and 1×1 convolutional layers. It contains in total 53 layers [18]. First, images are scaled to an input shape of 416×416 with 3 channels. After the feature extraction with Darknet53, the original image is converted into a feature map with a size 13×13. These feature maps are combined again to make two additional feature maps with sizes of 26×26 and 52×52. In other words, detection is performed on three levels, such that the feature map is transmitted to the two adjacent scales using up-sampling twice. For the first level, the high-resolution and low-level features are obtained. For the second level, the features are the combination of the $2\times$ up-sampled features from the first level and the features from the earlier layer via a residual skip connection. Similarly, for the third level, the low-resolution and high-level features are the combination of the $2\times$ up-sampled features from the second level and the earlier layer. On each feature map, each cell predicts three bounding boxes by means of three anchor boxes, finally selecting the most suitable bounding boxes. Three scales were selected to the targets of different sizes, which can now detect different sizes of targets. This is depicted in Figure 1.

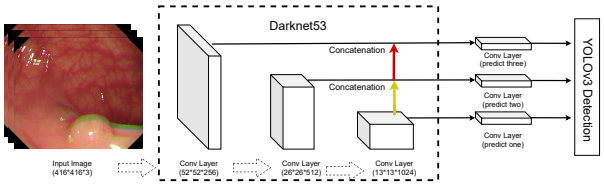

Fig. 1. Visualization of YOLOv3 Structure; red and yellow lines represent two-fold up-sampling.

The YOLOv3 was pre-trained with Common Object in Context (COCO) Image collection with over 118000 images [25]. The input image size is (416,416,3). Data augmentation is used. In order to balance the dataset, only polyps with fewer images (less than 250) were augmented. The augmentation strategies includes shifting, rotation, vertical or horizontal flipping, distortion, color jittering and different noises (including Gaussian noise, speckle and pepper&salt). In order to improve the performance of the model, 393 healthy images without polyps are added and they are not augmented. We used Adam [26] as the optimizer. The initial learning rate is 1×10^{-4}. The learning rate went down during training process to 1×10^{-8} we set L2 normalization for each layer. Early stopping was applied and patience equals 10 epochs.

E. Statistical analysis

We adopted one commonly accepted statistical method for evaluation the algorithm [27]. If the prediction of algorithm is

on a ground-truth polyp, then it is a true positive (TP) and only one positive case will be taken into consideration no matter how many predictions fall on the same polyp. The absence of a positive detection on an actual polyp is considered as one false negative (FN). If there is any detection label on a polyp-absence area, it is counted as false positive (FP). The per-image-sensitivity (S) or recall is defined as $TP/(TP + FN)$, precision (P) or positive predictive power is defined as $TP/(TP + FP)$. We also make use of the F1 score, defined as:

$$\frac{2 * (S * P)}{S + P},$$

and F2 score, defined as:

$$\frac{5 * (S * P)}{S + 4 * P}.$$

For evaluation, the sensitivity and precision of the model can be different depending on the confidence threshold to further adjust region of interests with various objectiveness score. Here, we adopted threshold 0.3 as used by Wittenberg et al. [28].

III. Results and Discussion

The test results for CVC-clinic database and ETIS-Larib Database are in TABLE III and TABLE IV, where our work is compared against the current state-of-the-art [10]. Figures 2 and 4 show examples of polyps including different sizes and morphology that were successfully detected. Figures 3 and 5 show examples of polyps including different sizes and morphology that were not successfully detected.

TABLE III
Comparison of Polyp Detection Performances on CVC-clinic Database

Studies	Model	TP (n)	FN (n)	FP (n)	S (%)	P (%)	F1 (%)	F2 (%)
Ours	YOLOv3	**591**	**55**	**21**	**91.5**	**96.6**	**94**	**93**
Wang et al. 2018 [29]	SegNet	570	76	42	88.2	93.1	91	89

TABLE IV
Comparison of Polyp Detection Performances on ETIS-Larib Database

Studies	Model	TP (n)	FN (n)	FP (n)	S (%)	P (%)	F1 (%)	F2 (%)
Ours	YOLOv3	**180**	**28**	**11**	86.5	**94.2**	**90.2**	**88.0**
Ahmad et al. 2019 [30]	-	-	-	-	**91.6**	75.3	88	**88**
Shin Y. et al. 2018 [31]	Inception Resnet	167	41	26	80.3	86.5	82	82
Qadir et al. 2021 [32]	MDeNetplus	**180**	**28**	28	86.5	86.1	86.3	86.5
Liu et al. 2019 [20]	YOLOV3	120	88	37	57.7	76.4	65.8	60.7

For CVC-Clinic database, even though the resolution lower than the training dataset, we showed both higher sensitivity and precision than the state-of-the-art. The resolution of feeding images is not an issue for our algorithm, which is the case for Wang et al. [29]. For Etis-Larib database, our model exhibits better precision than the study of Qadir et al. [32] and higher precision but lower sensitivity than Ahmad et al. [30]. It can be the case that high precision comes with the cost of low sensitivity therefore the F1 score becomes a necessary metric which combines sensitivity and precision. For both datasets, our F1 scores are better than the state-of-the-art. The detection results of study from Liu et al. [20] in Table IV, which also used the pretrained YOLOv3 model, are significantly lower than ours. There are several reasons to explain our improved performance. First, we have a large training dataset, which contains more images than other studies [20], [28], [29]. Second, we also augmented more than 10000

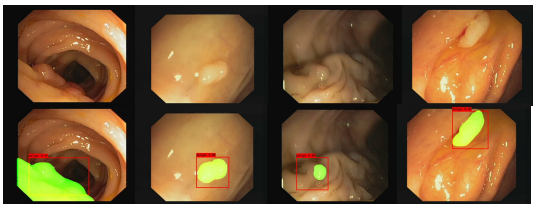

Fig. 2. True Positive cases of CVC-clinic DB. The red bounding box represents the prediction from our algorithm; the green area is the ground Truth.

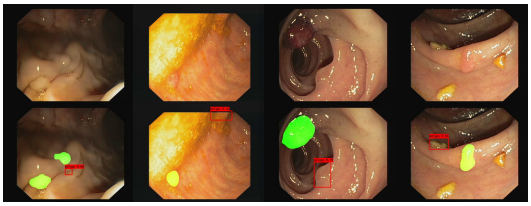

Fig. 3. False Positive and False Negative cases of CVC-clinic DB.The red bounding box represents the prediction from our algorithm; the green area is the ground Truth.

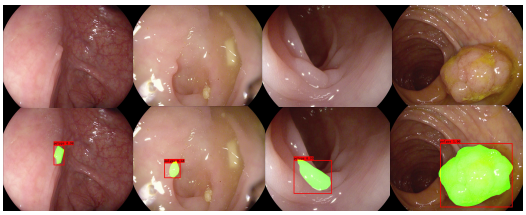

Fig. 4. True Positive cases of Etis-Larib DB.The red bounding box represents the prediction from our algorithm; the green area is the ground Truth.

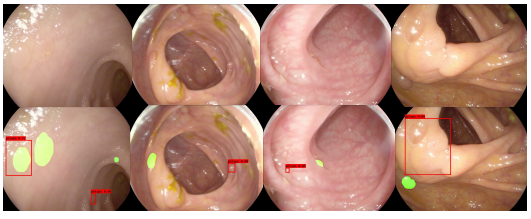

Fig. 5. False Positive and False Negative cases of Etis-Larib DB.The red bounding box represents the prediction from our algorithm; the green area is the ground Truth.

images with multiple augmentation methods and data balance strategy without uniformly augmenting all the polyps. Third, we enlarged our input size to 416×416 instead of 192×192 in Misawa, Masashi, et al [21]. Larger input size can feed more necessary information into the network. Fourth, our model was pre-trained with the COCO dataset described previously. In addition, our use of the Adam optimizer [26] instead of stochastic gradient descent and our use of a decaying learning rate could also be factors allowing us to achieve a better-performing model.

However, the system does have a few limitations. First, the algorithm is restricted to detect polyps in colonoscopy images but not taught to detect lesions outside the colon or in other examination formats. Second, the algorithm was trained to discriminate between normal mucosa and colonic polyps, but it is difficult to identify other intestinal content, see Figures 3 and 5. Third, the algorithm may miss small, flat and distant polyps. It is worthwhile to collect a large test dataset with various polyps and one independent healthy dataset given that most lumen are healthy in real clinical settings. In general, a more representative test dataset would be beneficial.

IV. Conclusion

We have presented an automatic polyp-detection algorithm based on YOLOv3. The detector has shown better performance than the current state-of-the-art. The results indicate that the ability of the algorithm to track polyps may be comparable to that of a skilled endoscopist. The high per-image-sensitivity could provide endoscopists with valuable visual assistance. Meanwhile, high precision is necessary to filter out false positive cases for endoscopists. Our model demonstrates performance that will not only provide a useful clinical tool, but also a solid starting point for further transfer learning to other colonoscopy endpoints such as cancer regrowth detection.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 857894.

References

[1] D. A. Lieberman *et al.*, "Guidelines for colonoscopy surveillance after screening and polypectomy: A consensus update by the US multi-society task force on colorectal cancer," *Gastroenterology*, vol. 143, no. 3, pp. 844–857, sep 2012.

[2] L. C. Seeff *et al.*, "How many endoscopies are performed for colorectal cancer screening? results from CDC's survey of endoscopic capacity," *Gastroenterology*, vol. 127, no. 6, pp. 1670–1677, dec 2004.

[3] A. G. Zauber *et al.*, "Colonoscopic polypectomy and long-term prevention of colorectal-cancer deaths," *N Engl J Med*, vol. 366, no. 8, pp. 687–696, feb 2012.

[4] N. N. Baxter, "Association of colonoscopy and death from colorectal cancer," *Ann Intern Med*, vol. 150, no. 1, p. 1, jan 2009.

[5] "Cancer facts and figures 2016."

[6] H. Pohl and D. J. Robertson, "Colorectal cancers detected after colonoscopy frequently result from missed lesions," *Clinical Gastroenterology and Hepatology*, vol. 8, no. 10, pp. 858–864, oct 2010.

[7] G. Urban *et al.*, "Deep learning localizes and identifies polyps in real time with 96% accuracy in screening colonoscopy," *Gastroenterology*, vol. 155, no. 4, pp. 1069–1078.e8, oct 2018.

[8] J. C. van Rijn *et al.*, "Polyp miss rate determined by tandem colonoscopy: A systematic review," *Am J Gastroenterology*, vol. 101, no. 2, pp. 343–350, feb 2006.

[9] C. M. C. le Clercq *et al.*, "Postcolonoscopy colorectal cancers are preventable: a population-based study," *Gut*, vol. 63, no. 6, pp. 957–963, jun 2013.

[10] A. Nogueira-Rodríguez *et al.*, "Deep neural networks approaches for detecting and classifying colorectal polyps," *Neurocomputing*, vol. 423, pp. 721–734, jan 2021.

[11] Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, may 2015.

[12] M. Raghu *et al.*, "Transfusion: Understanding transfer learning for medical imaging," in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds., vol. 32. Curran Associates, Inc., 2019.

[13] B. Q. Huynh *et al.*, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag*, vol. 3, no. 3, p. 034501, aug 2016.

[14] Y. Mori *et al.*, "Computer-aided diagnosis for colonoscopy," *Endoscopy*, vol. 49, no. 08, pp. 813–819, may 2017.

[15] Y. Wang *et al.*, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1379–1389, jul 2014.

[16] P. Brandao *et al.*, "Towards a computed-aided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks," *J. Med. Robot. Res.*, vol. 03, no. 02, p. 1840002, mar 2018.

[17] M. J. M. van der Valk *et al.*, "Long-term outcomes of clinical complete responders after neoadjuvant treatment for rectal cancer in the international watch & wait database (IWWD): an international multicentre registry study," *The Lancet*, vol. 391, no. 10139, pp. 2537–2545, jun 2018.

[18] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018.

[19] K. Jiang *et al.*, "Dense-layer-based YOLO-v3 for detection and localization of colon perforations," in *Medical Imaging 2021: Computer-Aided Diagnosis*, K. Drukker and M. A. Mazurowski, Eds. SPIE, feb 2021.

[20] M. Liu *et al.*, "Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network," *IEEE Access*, vol. 7, pp. 75 058–75 066, 2019.

[21] M. Misawa *et al.*, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal Endoscopy*, vol. 93, no. 4, pp. 960–967.e3, apr 2021.

[22] J. Bernal *et al.*, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, jul 2015.

[23] J. Silva *et al.*, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int J CARS*, vol. 9, no. 2, pp. 283–293, sep 2013.

[24] J. Redmon, "Darknet: Open source neural networks in c," 2013.

[25] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 740–755.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[27] A. I. Bandos *et al.*, "Area under the free-response ROC curve (FROC) and a related summary index," *Biometrics*, vol. 65, no. 1, pp. 247–256, may 2008.

[28] T. Wittenberg *et al.*, "Computer aided detection of polyps in whitelight-colonoscopy images using deep neural networks," *Current Directions in Biomedical Engineering*, vol. 5, no. 1, pp. 231–234, Sep. 2019.

[29] P. Wang *et al.*, "Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy," *Nat Biomed Eng*, vol. 2, no. 10, pp. 741–748, oct 2018.

[30] O. F. Ahmad *et al.*, "Tu1991 artificial intelligence for real-time polyp localisation in colonoscopy withdrawal videos," *Gastrointestinal Endoscopy*, vol. 89, no. 6, Supplement, p. AB647, 2019, dDW 2019 ASGE Program and Abstracts.

[31] Y. Shin *et al.*, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40 950–40 962, 2018.

[32] H. A. Qadir *et al.*, "Toward real-time polyp detection using fully CNNs for 2d gaussian shapes prediction," *Medical Image Analysis*, vol. 68, p. 101897, Feb. 2021.