Pattern Recognition 83 (2018) 209–219

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker

Ruikai Zhang[a,1], Yali Zheng[a,1], Carmen C.Y. Poon[a,*], Dinggang Shen[b,c,**], James Y.W. Lau[a]

[a] *Department of Surgery, The Chinese University of Hong Kong, Hong Kong*
[b] *Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina, Chapel Hill, NC, USA*
[c] *Department of Brain and Cognitive Engineering, Korea University, Seoul, South Korea*

ARTICLE INFO

Article history:
Received 15 August 2017
Revised 7 May 2018
Accepted 21 May 2018
Available online 30 May 2018

Keywords:
Smart cancer screening
Therapeutic endoscopy
Endoscopic Informatics
Body Sensor Network
Deep Learning
Health Informatics

ABSTRACT

A computer-aided detection (CAD) tool for locating and detecting polyps can help reduce the chance of missing polyps during colonoscopy. Nevertheless, state-of-the-art algorithms were either computationally complex or suffered from low sensitivity and therefore unsuitable to be used in real clinical setting. In this paper, a novel regression-based Convolutional Neural Network (CNN) pipeline is presented for polyp detection during colonoscopy. The proposed pipeline was constructed in two parts: 1) to learn the spatial features of colorectal polyps, a fast object detection algorithm named ResYOLO was pre-trained with a large non-medical image database and further fine-tuned with colonoscopic images extracted from videos; and 2) temporal information was incorporated via a tracker named Efficient Convolution Operators (ECO) for refining the detection results given by ResYOLO. Evaluated on 17,574 frames extracted from 18 endoscopic videos of the AsuMayoDB, the proposed method was able to detect frames with polyps with a precision of 88.6%, recall of 71.6% and processing speed of 6.5 frames per second, i.e. the method can accurately locate polyps in more frames and at a faster speed compared to existing methods. In conclusion, the proposed method has great potential to be used to assist endoscopists in tracking polyps during colonoscopy.

1. Introduction

Colorectal cancer (CRC) is the fourth cause of cancer death worldwide, with around 1,360,000 newly diagnosed cases and 694,000 people dying from the disease each year [1]. Colonoscopy is one of the most effective tools for identifying and removing polyps before they developed into CRC and is estimated to contribute to a 25% decline in mortality for CRC [2]. Moreover, when equipped with advanced techniques such as narrow band imaging, endoscopists can phenotype the disease in vivo and in situ by predicting its histological characteristics during colonoscopy. Nevertheless, colonoscopy is a highly operator-dependent procedure and up to 25–28% of polyps can be missed in a single colonoscopy [3,4].

2. Related works

Different Computer-Aided Detection (CAD) methods for polyp detection in colonoscopy have been actively investigated, but few of them were evaluated in clinical setting [5]. Compared to the CAD tasks in non-medical domain, automated polyp detection is technically challenging in practice since 1) the same type of polyp can vary greatly in size, color and texture, and 2) many polyps do not stand out clearly from the surrounding mucosa. Previous CAD systems usually extracted handcrafted features from polyp images, such as their color and texture [6,7], shape or appearance [8–11], or the combination of these features [12,13], and trained a classifier to distinguish the polyp from the surroundings. The sensitivity of such method can vary largely between 48 and 88% in unaltered videos and selected frames, respectively [12].

Deep learning has revolutionized the paradigm of machine learning and achieved significantly better performance in various classification tasks [14], including analyzing medical images for disease diagnosis and treatment like lung cancer detection [15,16], skin cancer classification [17], colorectal cancer diagnosis [18], breast cancer classification [19] and cell based therapy [20]. It is intuitive to use deep learning methods to automatically learn abstract and discriminative features from medical data to

[*] Corresponding author.
[**] Corresponding author at: Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina, Chapel Hill, NC, USA.

E-mail addresses: rzhang@surgery.cuhk.edu.hk (R. Zhang), ylzheng@surgery.cuhk.edu.hk (Y. Zheng), cpoon@surgery.cuhk.edu.hk (C.C.Y. Poon), dgshen@med.unc.edu (D. Shen), laujyw@surgery.cuhk.edu.hk (J.Y.W. Lau).

[1] Authors contribute equally to this work.

https://doi.org/10.1016/j.patcog.2018.05.026
0031-3203/© 2018 Elsevier Ltd. All rights reserved.

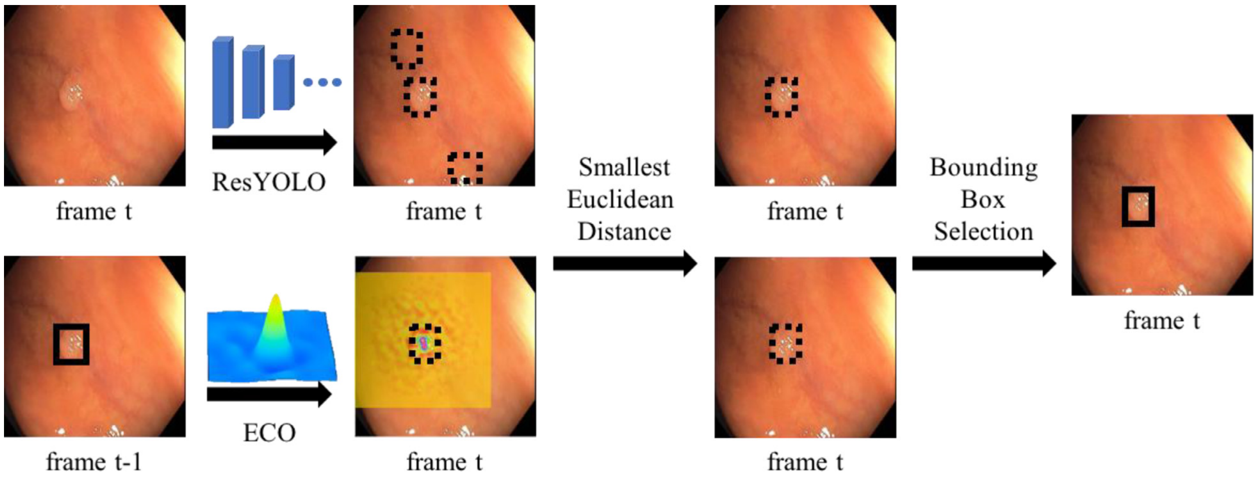

Fig. 1. Pipeline of the proposed computer-aided detection algorithm RYCO.

overcome the difficulties of designing robust handcrafted features based on the domain knowledge. The 2-Dimensional (2D) Convolutional Neural Network (CNN) architecture was adopted by most existing studies for polyp detection [21–25], of which the detection task was formulated as a classification problem. The typical strategy is to first train a 2D CNN for classifying pre-selected candidate patches as polyp or non-polyp, and then use a sliding window strategy and summarize the predictions on patches to locate polyps in a test image [21]. The network is however complex and relatively slow since the whole image has to be searched by the sliding window strategy. Moreover, each image is processed independently and the temporal information between different frames is not fully utilized. A recent study [26] attempted to use 3D CNN for polyp detection by leveraging spatial-temporal information from colonoscopy videos using a single output probability map. The method, however, requires 1.23 seconds and 8 frames latency to process a single frame and therefore is difficult to be used in real clinical setting.

In order to search for an object efficiently in a frame, a regression-based CNN named You Only Look Once (YOLO) has been recently proposed to predict the object bounding boxes and class probability of the object in one-shot [27]. Unlike the classification-based detectors, encoding contextual information based on the whole image makes YOLO extremely fast.

In this paper, a novel CAD method called RYCO was proposed for fast polyp detection in endoscopic videos. RYCO learned spatial features by an extension of YOLO which we addressed as ResYOLO detector in the rest of this paper, and integrated temporal information by a Discriminative Correlation Filter (DCF) based tracking method called ECO tracker [28]. The contributions of our proposed method are as follows: To our best knowledge, this is the first study 1) to use regression-based deep CNN methods with residual learning for polyp detection in order to achieve state-of-the-art performance at a fast speed, and 2) to incorporate in the model temporal information using DCF-based method to accurately locate polyps in more frames. It is expected that the proposed method will have great potential in cancer prevention.

3. Method

In order to accurately and efficiently detect polyps in colonoscopy videos, the proposed CAD method (RYCO) learns spatial information of current frame and also incorporate temporal information from previous frames. Fig. 1 shows the pipeline of the proposed method, where a regression-based deep CNN with residual learning (ResYOLO) detection model is built to locate a polyp in a frame and decide whether the tracked polyp is still in the view. Meanwhile, assuming that a polyp will not jump from one location to another between two consecutive frames, a tracker is introduced to provide a stable guidance on the polyp location to ensure robustness of the detection to jittering of frame quality.

3.1. Spatial feature learning

3.1.1. Regression-based 2D CNN (YOLO)

Classification-based detection architecture like region-proposal-based 2D CNN usually requires two stages to detect: 1) to generate potential bounding boxes and 2) to run classifier on the proposed boxes. Unlike the classification-based detection architecture, the regression-based detection architecture, such as YOLO, only needs a single CNN to simultaneously predict multiple bounding boxes and their class probabilities. Therefore, YOLO can be at least 2.5 times faster than a region-proposal-based 2D CNN model with similar performance [27]. In addition, regression-based CNN encodes information of the whole image instead of regions to make predictions and therefore, it is less vulnerable to background errors [27].

The pipeline of YOLO was to first divide the input image into $G \times G$ grids, and a 2D CNN architecture was trained to predict B bounding boxes for each grid. Each grid was assigned a conditional class probability P for C classes if the grid contains an object. A total of 5 indicators were used to describe each bounding box including center coordinates (x, y), width and height of the box (w, h) and confidence score *Conf*. Thus, each grid was described by a vector of K elements, where $K = 5 \times B + C$. The output of candidate bounding boxes for the whole image was a matrix of size of $G \times G \times K$. Finally, non-maximum suppression was used to remove redundant candidate bounding boxes. The loss function for optimizing the regression model was comprised of loss for grids labeled as object and those as no-object described in the following equation:

$$L = \sum_{G^2}\left(\sum_{B} O * \left(\lambda_{bbox} * f\left(x, y, \sqrt{w}, \sqrt{h}\right) + \lambda_{obj} * f(Conf)\right)\right.$$
$$\left. + (1-O) * \lambda_{noobj} * f(Conf) + \sum_{C} O * f(P)\right). \qquad (1)$$

The higher the intersection over union the predicted bounding box is with ground truth, the higher the *Conf* is. $f(\cdot)$ denotes the square of the differences between the variable and its ground truth. O equals 1 if any object is within the selected grid, and vice

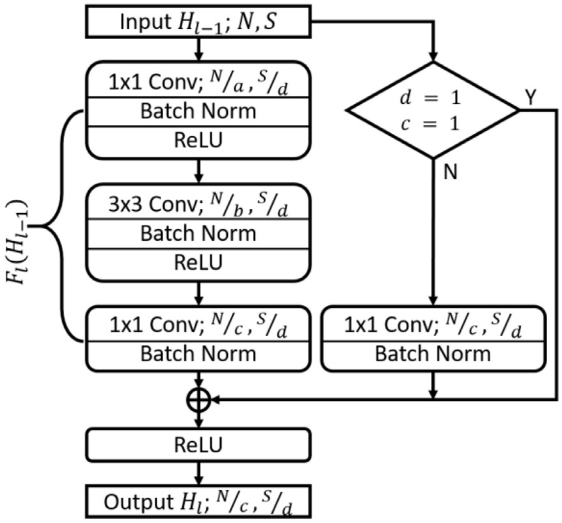
Input H_{l-1}; N, S
1x1 Conv; $^N/_a$, $^S/_d$
Batch Norm
ReLU
3x3 Conv; $^N/_b$, $^S/_d$
Batch Norm
ReLU
1x1 Conv; $^N/_c$, $^S/_d$
Batch Norm
$F_l(H_{l-1})$
$d = 1$
$c = 1$
Y
N
1x1 Conv; $^N/_c$, $^S/_d$
Batch Norm
ReLU
Output H_l; $^N/_c$, $^S/_d$

Fig. 2. Residual learning module of our proposed ResYOLO detector.

versa. To focus on optimizing loss of grids with objects, different weights λ were added to different terms, where λ_{bbox}, λ_{obj} and λ_{noobj} were set to 5, 1 and 0.1, respectively. The different weights λ were selected according to [27].

3.1.2. Residual learning module

Different from detecting objects in natural images, where the objects usually stand out from their background, a polyp usually looks very similar to its surrounding mucosa. In order to extract features that better describe endoscopic and histological features of different polyps, we proposed to introduce residual learning modules in the YOLO CNN architecture for feature learning [29]. Fig. 2 shows the residual learning modules of our proposed detector architecture (ResYOLO). The rationale of including the residual learning modules in our design is as follows: it was reported that the depth of the CNN is crucial for performance as more layers the network has, the richer and more abstract features can be learned. Nevertheless, as the CNN architecture gets deeper, it becomes more difficult to train, especially when only limited training data samples are available. By introducing the residual module as a skip network structure in the design, an enrich set of features can be extracted by the proposed deep architecture even with limited training data.

ResYOLO used a residual mapping $F(\cdot)$ that learned by the following equation:

$$H_l = F_l(H_{l-1}) + I(H_{l-1}) \tag{2}$$

where $I(\cdot)$ is the identity function, l is the index of layer, and H_{l-1} and H_l are the input and output. Three convolutional blocks were used for residual feature learning, where *a, b, c* and *d* were coefficients to manipulate the number of channels N and dimension S of the data. If both the number of channels and the dimensions of H_{l-1} and H_l were the same, the identity function $I(\cdot)$ can be described as $I(x) = x$. While in some cases, coefficient c was set to 0.5 to increase the number of features to extract and d was set to 2 to down-sample instead of using pooling layers. To match up the size of $I(H_{l-1})$ with $F_l(H_{l-1})$, an additional 1x1 convolutional layer was added and the identity function $I(\cdot)$ was described as $I(x) = Wx$ instead.

3.1.3. Training procedure for ResYOLO

Since images extracted from endoscopic videos are subject to motion artifacts and unexpected reflections, two strategies were used to introduce variations in the training dataset. First, a set of images without a polyp were used as samples for a negative class during training, i.e. the number of classes C used in our model was 2. The additional negative class ensures: 1) the training dataset owns a more generalized data distribution; and 2) the trained model will be more sensitive to suspicious frames and reduce false positives. Second, video frames from the training dataset were augmented by rotation, up-down flip, and left-right flip, as well as by additive Gaussian smoothing and random contrast and brightness to mimic the motion blurs caused by bowel and endoscope movements and variations of light conditions during colonoscopy. The sequence and methods for augmentation were randomly selected during training. For example, some of the images were augmented by rotation, up-down flip, and then Gaussian smoothing, and some were augmented by all the 5 methods mentioned above.

Fig. 3 shows the details of ResYOLO detector architecture, where 16 residual learning modules were used and then followed by 3 convolutional layers as classifier. The 1st and 2nd numbers after each semicolon in Fig. 3 represents the number of channels and the dimension of the output, respectively. Images were resized to 448 × 448 × 3 pixels and subtracted with the mean of training dataset as input. The grid size G is set to be 14, which was also the dimension of the output matrix. Moreover, the number of candidate bounding boxes B predicted for each grid was set to 2 so that the number of channels of the output matrix was 12 $(5 * 2 + 2)$. The maximum number of objects that can be detected in this configuration was 196 (i.e., 14*14), which is assumed to be sufficient for polyp detection.

3.2. Integration of temporal information

3.2.1. DCF-based tracking method

Although the proposed ResYOLO detector performs well based on the learned spatial information, the lack of temporal information makes it vulnerable to unexpected jittering and artifacts. For example, different bounding boxes can be detected on similar consecutive frames. A 3D CNN model proposed to jointly learn spatial and temporal information; however, the drawback of the method is that it is computationally complex [26]. We proposed in this study to adopt a real-time online object tracker, called Efficient Convolution Operators (ECO) [28], to locate the polyp on the current frame based on what was detected from the previous frames. ECO introduced a factorized convolution operator to learn a dimension-reduced multi-channel convolution filter f from samples tracked along previous frames, where the filter f transfers multi-resolution feature maps to a continuous spatial domain $t \in [0, T)$. Instead of learning the filter f based on all previously tracked samples, it learned on L selected Gaussian components resulted from these samples using Gaussian Mixture Model. The objective function for optimizing f is described as:

$$E(f) = \sum_L \pi \left\| S_f\{\mu\} - y \right\|_{L^2}^2 + \sum_D \|\omega f\|_{L^2}^2 \tag{3}$$

where μ was the Gaussian mean, π was the prior weight for each component, ω was the weight for penalty term, and $S_f\{\mu\}$ and y was the predicted and labeled detection score, respectively. The L^2-norm for $\|z\|_{L^2}^2$ can be described as $\|z\|_{L^2}^2 = 1/T \int_0^T |z(t)|^2 dt$. The output of the proposed ResYOLO detector on the first detected polyp was used as the initial input for ECO tracker.

3.2.2. Integration of ResYOLO detector and ECO tracker

During colonoscopy, a polyp can sometimes be occluded by saline water flushing or moved out of sight due to bowel or endoscope movements and reappear at a totally different location several frames later. When the polyp moves out of sight, the tracker needs to be informed. Therefore, the detected polyp given by ResY-

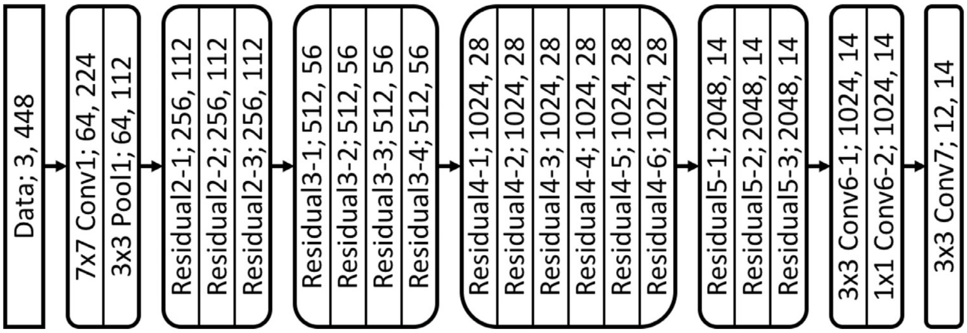

Fig. 3. Architecture of the proposed ResYOLO Detector.

Table 1
AsuMayoDB data split for experiment in comparison to MICCAI 2015 endoscopic vision challenge.

Challenge data split	Experiment data split	Num. of polyp images	Num. of non-polyp Images	Num. of videos
Training	Training	3237	11,295	16
	Validation	619	3751	4
Testing	Testing	4313	13,261	18

OLO detector will be used to initialize the tracker as well as to decide when to stop tracking.

Since non-polyp images were labeled as the negative class for training, only the candidate bounding boxes for polyp is used at this stage. At present, it is assumed that only one polyp will present in each frame. The candidate bounding box with the highest confidence score *Conf* in the frame and at the same time higher than a threshold *T* can be used to initialize tracking. Once the tracker was activated, the final decision for frame *t* was then decided by both ResYOLO detector and ECO tracker and each decision made afterwards was determined by the predictions from previous frames (see Fig. 1). The candidate polyp location given by ECO tracker was used as a rough location of polyp for frame *t*. The bounding box given by ResYOLO detector with the smallest Euclidean distance to the rough location given by ECO tracker was then taken as a second candidate bounding box. If the mean *Conf* of the previous *M* (including the one from current frame) frames, denoted as *Conf*′, was larger than the threshold *T*, then the current frame was suspicious of containing polyp; otherwise, stop tracking and re-initialize tracking when polyp was detected again on later frames. The *Conf*′ is calculated by:

$$Conf' = \frac{1}{M} * \sum_{m=0}^{M-1} Conf_{t-m}. \qquad (4)$$

If the $Conf_t$ of the selected candidate bounding box given by ResYOLO detector was higher than the said *Conf*′, then this bounding box will be considered as the final detection for current frame *t* and used as one of the samples for online training of ECO tracker.

4. Experiments

4.1. Experimental setup for video database

In this paper, the same training and testing datasets used in MICCAI 2015 Endoscopic Vision Challenge were used [30]. As shown in Table 1, 20 and 18 videos were used for training and testing in the challenge setup, respectively. Among these 20 videos used for training, 10 of them were videos without polyps and the other 10 have at least 1 frame with polyp. In this study, to have sufficient size of training data while avoid overfitting, these 20 videos were further divided into a training set and a validation set by the ratio of 4:1, where 4 of the training videos, i.e., 2 without polyp and 2 with polyp, were randomly selected to validate the ResYOLO model. The performance of ECO, ResYOLO and RYCO were tested on 18 videos, where 6 of them have at least 1 frame with polyp, and 3 of them have polyp in every frame. The experiment of using ECO was under the assumption that the ground truth of the polyp location of 1st frame of appearance was given.

All images were extracted from the videos, and non-informative frames were removed by a threshold. The rest of the images were resized to 448x448x3 pixels. The binary mask provided for each polyp image was transformed to the smallest bounding box that can cover the whole polyp as the ground truth. The training set was augmented to 25,600 for each class as previously described. The weights pre-trained from the 1000-class general object classification (ImageNet) [31] was used as initial weights before fine-tuning the proposed ResYOLO. The models were trained either (1) only by images with polyps, or (2) by images with or without polyps. Similar to the optimizing strategy utilized in YOLO [27], a lower learning rate 10^{-5} was used for the 1st epoch, followed by a higher learning rate 10^{-4} and gradually decreased by a factor of 10. The decay rate, momentum and batch size were set as 0.0005, 0.9 and 5, respectively. Threshold *T* for initializing tracking as well as determining suspicious frames was 0.2. The number of frames *M* for calculating *Conf*′ was 6. The performance of using different numbers of frames *M* for calculating the *Conf*′ was reported.

4.2. Evaluation metrics

Evaluation metrics used in MICCAI 2015 Endoscopic Vision Challenge [30] were used in this study:

1) True Positive (TP). When the centroid of the predicted box falls inside the ground-truth bounding box, the detection is considered as a TP. It is worth noting that, if the centroids of multiple predicted boxes fall inside the same ground-truth bounding box, it will only be counted as one TP.
2) False Positive (FP). When the centroid of the predicted box falls outside of the ground truth, it will be considered as a FP.
3) True Negative (TN). When no predicted boxes were given on a non-polyp image, it will be considered as a TN.
4) False Negative (FN). When none of the centroids of the predicted boxes fall inside any ground-truth bounding boxes, it will be considered as a FN.
5) $Specificity(Spec) = TP/(TP + TN)$.
6) $Precision(Prec) = TP/(TP + FP)$.

Table 2
Summary of results for all videos.

Method	TP	FP	TN	FN	Spec (%)	Prec (%)	Rec (%)	F1 (%)	F2 (%)
ASU	2636	184	13,149	1677	**98.6**	**93.5**	61.1	73.9	65.7
CUMED	3081	769	13,010	1232	94.4	80.0	71.4	75.5	73.0
3D CNN	3062	414	N/A	1251	N/A	88.1	71.0	78.6	73.9
ECO	3032	3104	N/A	1281	N/A	49.4	70.3	58.0	64.8
YOLO	2080	384	13,141	2233	97.2	84.4	48.2	61.4	52.7
ResYOLO	2779	282	13,153	1534	97.9	90.8	64.4	75.4	68.4
RYCO	3087	398	13,057	1226	97.0	88.6	**71.6**	**79.2**	**74.4**

Table 3
Summary of results for videos with at least one frame with polyp.

Method	TP	FP	TN	FN	Spec (%)	Prec (%)	Rec (%)	F1 (%)	F2 (%)
ASU	1218	92	1864	1431	**95.3**	**93.0**	46.0	61.5	51.2
CUMED	1439	600	1692	1210	73.8	70.6	**54.3**	61.4	56.9
3D CNN	1424	385	N/A	1225	N/A	78.7	53.8	63.9	57.4
ECO	1371	3101	N/A	1278	N/A	30.7	51.8	38.5	45.5
YOLO	932	371	1770	1717	82.7	71.5	35.2	47.2	39.2
ResYOLO	1192	218	1831	1457	89.4	84.5	45.0	58.7	49.6
RYCO	1433	358	1752	1216	83.0	80.0	54.1	**64.5**	**57.8**

Table 4
Summary of results for videos with all frames with polyp.

Method	TP	FP	FN	Prec (%)	Rec (%)	F1 (%)	F2 (%)
ASU	1418	40	246	97.2	85.2	90.8	87.4
CUMED	1642	149	22	91.7	98.7	95.0	97.2
3D CNN	1638	0	26	**100.0**	98.4	99.2	98.7
ECO	1661	3	3	99.8	**99.8**	**99.8**	**99.8**
YOLO	1457	148	207	90.8	87.6	89.1	88.2
ResYOLO	1587	48	77	97.1	95.4	96.2	95.7
RYCO	1654	7	10	99.6	99.4	99.5	99.4

7) $Recall\ Rate(Rec) = TP/(TP + FN)$.
8) $F1\ score = 2 * Prec * Rec/(Prec + Rec)$.
9) $F2\ score = 5 * Prec * Rec/(4 * Prec + Rec)$.
10) Frame per second (fps) for Processing time

The performance of the proposed method was compared with a state-of-the-art method (denoted as 3D CNN) [26], the methods with top 2 performance during MICCAI 2015 Endoscopic Vision Challenge (ASU and CUMED) and the baseline model YOLO. For fair comparison, YOLO was trained with the same dataset described above. Note that the TN was not reported in 3D CNN [26] and was not tested for ECO since only videos with polyps were evaluated on ECO. The 95% confidence interval (CI) for the proposed RYCO method was calculated based on the binomial equation, i.e.

$$\mathrm{o} \pm 1.96 * \sqrt{\frac{\mathrm{o} * (1 - \mathrm{o})}{\mathrm{n}}}, \tag{5}$$

where o is the observation and n is the number of experiment samples.

4.3. Experiment results

The overall results on all 18 testing videos were summarized in Table 2. Our proposed method RYCO achieved the following performance: Specificity = 97.0% (95% CI:96.7%–97.3%), Precision = 88.6% (95% CI:87.5%–89.7%), Recall rate = 71.6% (95% CI:70.3%–73.0%), F1 score = 79.2% (95% CI:78.0%–80.4%), and F2 score = 74.4% (95% CI:73.2%–75.7%). The average speed of ResYOLO and RYCO are 14.0 and 6.5 fps, respectively, when running on a standard PC with a 3.3 GHz Intel(R) Xeon(R) E3-1226 CPU and an NVIDIA GeForce GTX TITAN X GPU. Tables 3 and 4 show the analysis for videos with at least one frame with polyp and videos with all frames with polyp, respectively. Table 5 shows in a stepwise manner the improvement in performance of each strategy mentioned in Section 3, using $\mathrm{YOLO_o}$ as the baseline model. The performance curves of different evaluation criteria for choice of different numbers of frames M for calculating the $Conf'$ were illustrated in Fig. 4.

5. Discussion

Studies in the epidemiology and molecular pathology of CRC [32,33] are emerging into a new field of study, i.e. molecular pathological epidemiology [34,35], with an aim to better understand and prevent CRC carcinogenesis. Colonoscopic polypectomy has been considered as the gold standard for detecting and resecting colorectal lesions before they develop into cancer. Several practical factors will influence the polyp recognition rate in a colonoscopy, including characteristics of the patient, bowel preparation, endoscopists' experience and properties of the polyp. Polyps that are small in size (<10 mm) [36], flat shape, left-side located [3], and sessile serrated in type are much easier to be missed.

CAD has gained great attention for its potential to improve polyp detection rate by highlighting possible lesions during the colonoscopy procedure [5]. This study proposed a novel CAD method RYCO for fast polyp detection and demonstrated it performed better than state-of-the-art algorithms when evaluated on a public colonoscopic video database. Table 2 shows that RYCO can accurately locate more polyps compared to existing algorithms. ASU achieves the best precision amongst all methods presented in this paper; however, this is at the expense of trading off recall rate. When comparing ASU with the proposed RYCO, RYCO outperforms ASU in terms of both F1 and F2. In practice, due to unexpected bowel or endoscope movements, polyps can easily move out of sight, which is a situation similar to that presented in Table 3. In this situation, compared to all other methods, RYCO maintains a relatively low false prediction and high accuracy in locating polyps (F1-score = 64.5% and F2-score = 57.8%).

Compared to previous classification-based polyp detectors, the ResYOLO architecture used in RYCO avoids generating ROIs by additional region proposal methods and predicts polyp locations directly on the whole image. The proposed RYCO has less frame latency than the 3D CNN model. RYCO incorporated temporal information in two aspects: 1) ECO utilized the features learned from L previous frames to make decision at the current frame; and 2) when integrating ResYOLO with ECO, the confidence scores of the predicted boxes of the current frame were averaged over the cur-

Table 5
Performance improvement by each proposed strategy.

	$YOLO_o$					ResYOLO	RYCO
+ fully convolutional layers		√	√	√	√	√	√
+residual module			√	√	√	√	√
+larger grid size *G*				√	√	√	√
+additional augmentation					√	√	√
+non-polyp images						√	√
+ECO tracker							√
F1 (%)	60.0	60.9	69.6	71.8	72.2	75.4	79.2
Improve by (%)	0.0	0.9	8.7	2.2	0.4	3.2	3.8

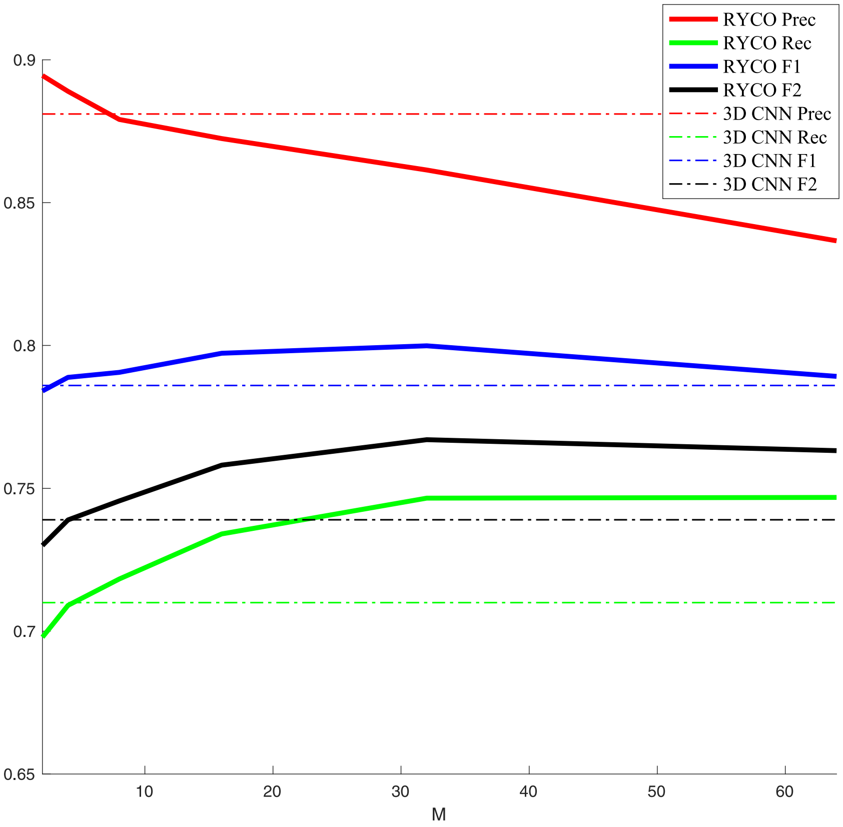

Fig. 4. Performance curve of using different numbers of frames *M* for calculating the *Conf′*.

rent frame and *M*-1 previous frames to arrive at the final prediction confidence score. We did not directly embed the temporal information in the ResYOLO detector because a 3D CNN model is very time consuming and inapplicable for real-time polyp detection, as shown in [26]. The results in Table 2 showed that our proposed method had similar, if not better, performance than a complex 3D CNN model that was jointly trained for spatial and temporal information. Moreover, RYCO can be run at a speed of 6.5 fps which is about 7 times faster than the complex 3D CNN model (0.8 fps) under similar hardware setup.

As illustrated in Table 5, incorporating the residual learning modules in the proposed design contributes mainly to the improvement in accuracy, comparing to the baseline model $YOLO_o$. Augmentation by additive Gaussian smoothing and random contrast and brightness does not have significant improvement in F1 score; however, it helps reduce FP from 809 to 692. By introducing a negative class for spatial feature learning, the model was able to better discriminate suspicious images and thereby raising the F1 score by 3.2% when compared to training the model with polyp images only. Compared to $YOLO_o$, ResYOLO and RYCO have improved by 15.4 and 19.2% in terms of F1 score, respectively.

The inclusion of the confidence score *Conf* from the bounding box detected in the previous frame was based on the assumption that two consecutive frames do not possess sudden changes. If a polyp was detected in previous frames, it is highly likely to appear in the next frame. If at frame $t-1$, another two unexpected false positives occurred, RYCO may not be able to remove these two false positives in next frame *t*; however, there is a high possibility that RYCO can remove these false positives in the subsequence frame, $t+1$. The performance of RYCO can be further tuned by the parameter *M* to arrive at a different score *Conf′*. As illustrated in Fig. 4, a relatively high *M* can help the model to correctly locate polyps in more images, but at an expense of increasing false positives.

ECO tracker achieved a good performance on videos with only polyp images, i.e. it performs well when a polyp was correctly located in the 1st frame. Nevertheless, using only the ECO tracker performs poorly when the polyps to be tracked were occluded or moved out of sight due to large endoscope or bowel move-

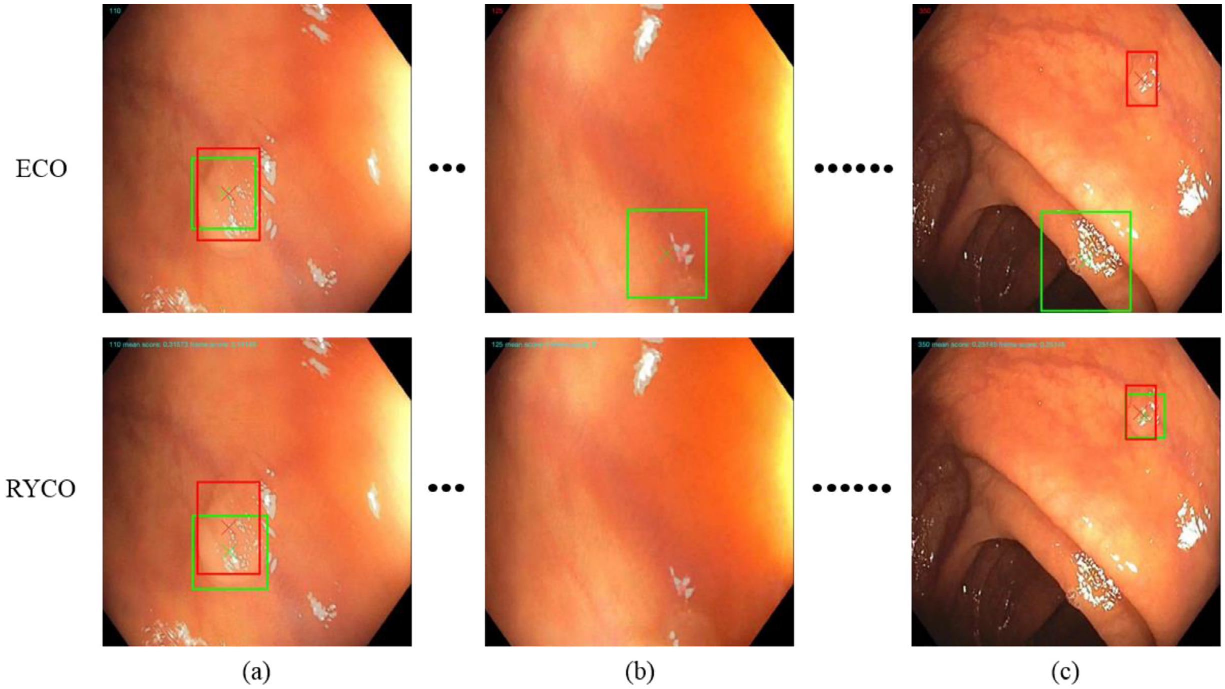

Fig. 5. Typical prediction differences between ECO and RYCO. Images were selected from AsuMayoDB with labeled ground truth (red boxes) as well as prediction by ECO and RYCO (green boxes): column (a) 110th frame, column (b) 125th frame and column (c) 350th frame of testing video 11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

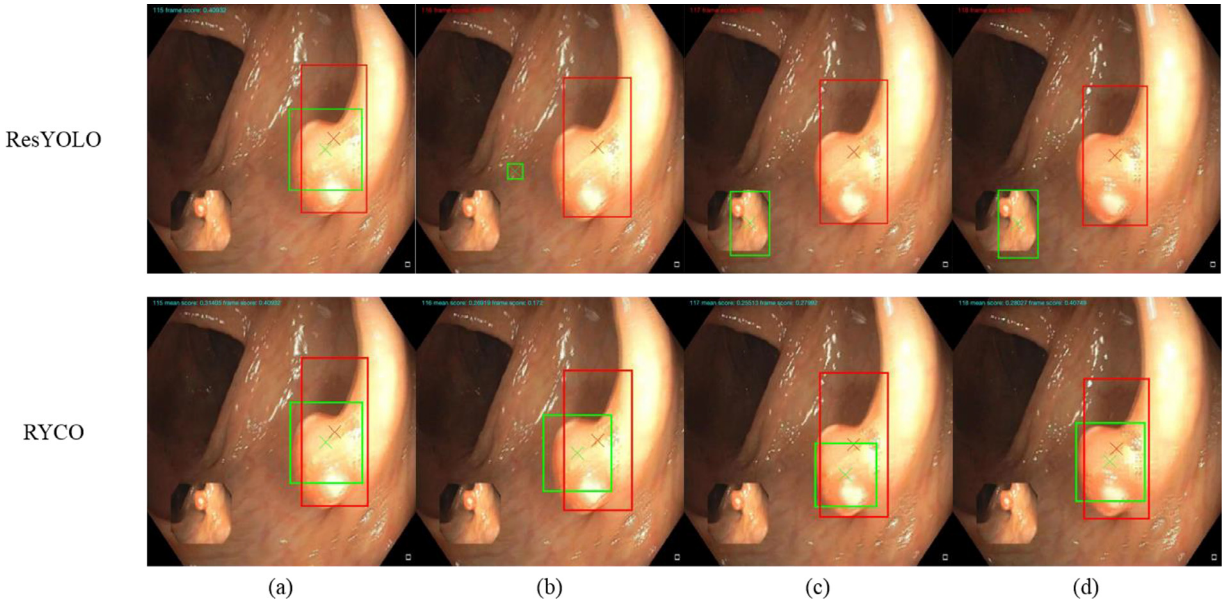

Fig. 6. Typical prediction differences between ResYOLO and RYCO. Images were selected from AsuMayoDB with labeled ground truth (red boxes) as well as prediction by ResYOLO and RYCO (green boxes): column (a) 115th frame, column (b) 116th frame, column (c) 117th· and column (d) 118th frame of testing video 13. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

ments, as shown in Fig. 5. On the other hand, if relying only on ResYOLO detector, some polyps can be miss-detected due to unexpected artefacts or jittering even when the polyp in previous frame was correctly detected, as shown in Fig. 6. Therefore, the proposed RYCO combines ResYOLO and ECO in the pipeline in order to achieve a high precision while maintaining a reasonable recall rate.

Fig. 7 summarizes typical errors made by RYCO when used on colonoscopy with adequate bowel preparation in this study. As shown in Fig. 7(a)–(c), RYCO failed to detect polyps when their patterns were not typical, or when they did not stand out from the surrounding mucosa. On the other hand, the ridge of mucosa can sometimes trigger a false alarm, as illustrated by Fig. 7(d). In other situation, the bounding box predicted by RYCO was not completely within the labelled ground truth, as shown in Fig. 7(e), and therefore has been counted towards a false positive according to the definition of the Challenge. Nevertheless, the prediction still serves as a good guidance of the location of the polyp. Lastly, judging from the ground truth labeled for its neighboring frames, it

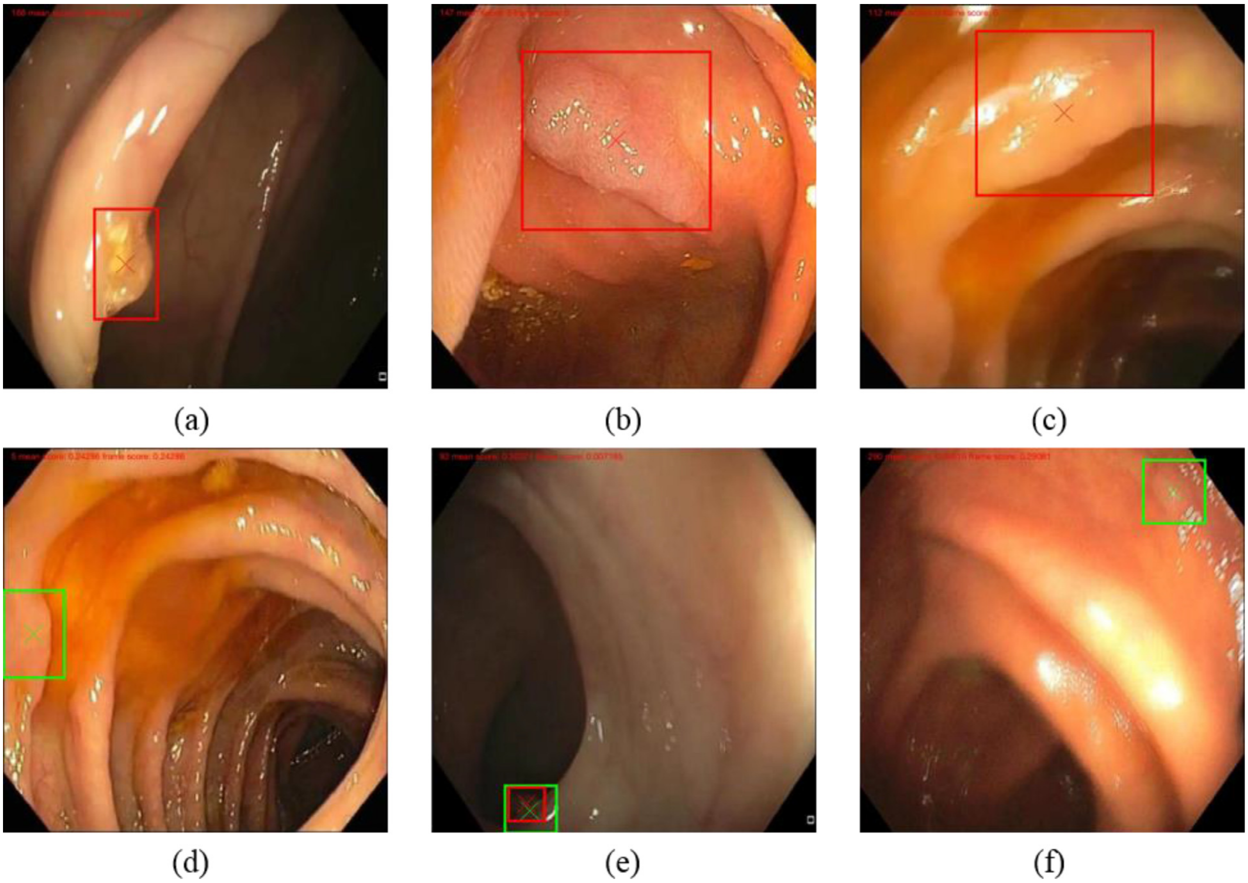

Fig. 7. Typical false negatives (a)–(c) and false positives (d)–(f) made by RYCO during colonoscopy with adequate bowel preparation. Images were selected from AsuMayoDB with labeled ground truth (red boxes) and prediction by RYCO (green boxes): (a) 158th frame of testing video 5, (b) 147th frame of testing video 7, (c) 112th frame of testing video 8, (d) 5th frame of testing video 8, (e) 92nd frame of testing video 10, and (f) 290th frame of testing video 11. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

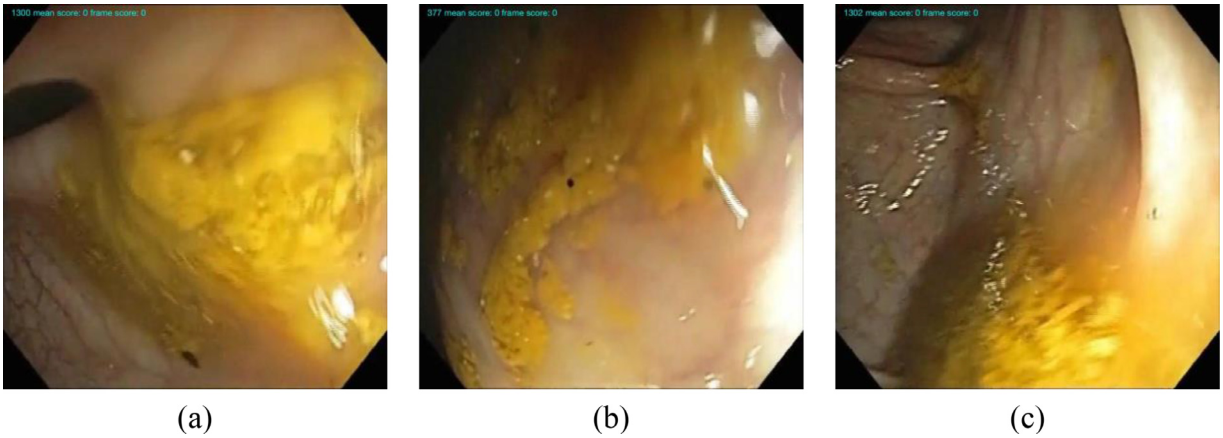
1300 mean score: 0 frame score: 0
377 mean score: 0 frame score: 0
1302 mean score: 0 frame score: 0
(a)
(b)
(c)

Fig. 8. Typical true negatives made by RYCO during Colonoscopy with inadequate bowel preparation. Images were selected from AsuMayoDB with prediction by RYCO (green boxes): (a) 1300th frame of testing video 15, (b) 377th frame of testing video 16, (c) 1302nd frame of testing video 17. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

was also arguable that some of the false positives were contributed by mislabeled ground truth as shown in Fig. 7(f).

The evaluation database consists of a few colonoscopy videos that were taken under inadequate bowel preparation and Fig. 8 shows the typical performance of RYCO under this situation. Although the bowel was filled with feces, RYCO did not mis-detect those regions as polyps. Enlarging the database can further help to evaluate whether the proposed algorithm can handle unexpected situations like these.

Labeled data has become the main bottleneck of current data-driven feature representation techniques, especially in the medical domain, where data requires more time and domain knowledge to label [37]. Although the evaluation dataset used in this study only contains 18 colonoscopy videos, it is one of the largest, frame-by-frame-labeled colonoscopy dataset publicly available. The dataset consists of 17,574 frames, which can be considered at present a sizable dataset in this medical domain. In future, further evaluation of the proposed method to cover more different scenarios in real-life colonoscopy will be of interest. First, instead of short video clips of 2 min, full colonoscopy procedure over 10–30 min should be recorded for evaluation. Second, situations such as those with severe inadequate bowel preparation are currently lacking. Third, all video clips used in this study are acquired under white light. While narrowband imaging is now becoming more frequently used in colonoscopy, the data distribution and polyp features from these images are very different and should be trained separately. Fourth, the current database did not provide information on the pathology of the lesions in the videos. It is therefore difficult to know whether it covers some rare cases of colorectal lesions such as sessile serrated polyp, which is more likely to be missed even by experienced endoscopists.

Limited by the current evaluation database, the performance of the proposed algorithm in real-life colonoscopy is difficult to be conclusive at this stage. Despite of these limitations, this study demonstrated the potential of RYCO for fast and accurate computer-aided polyp detection during colonoscopy compared to other state-of-the-art algorithms.

6. Conclusion

In conclusion, the proposed RYCO pipeline detects polyps in the AsuMayoDB with a precision of 88.6% and recall of 71.6% at a processing speed of 6.5 frames per second. The proposed method can accurately locate polyps in more frames and at a faster speed, compared to previous algorithms. In future, we intend to extend RYCO by working on a less computationally complex CNN architecture to achieve real-time processing speed, while maintaining a high sensitivity in accurately locating polyps. Moreover, back-to-back clinical study will be performed to evaluate the potential of increasing polyp detection rate as well as shortening withdrawal time. The algorithm will be useful to the emerging therapeutic wireless endoscopic capsules [38], or to integrate with classifier that judges whether a polyp should be resected or not [15]. A prospective randomized control trial can further verify the effectiveness of the proposed pipeline in assisting endoscopists to improve adenomatous detection rates or to decide whether a polyp should be resected or not during colonoscopy.

Acknowledgment

This project is supported in part by General Research Fund (GRF/14202417) and Innovation and Technology Fund (ITF/337/16FP).

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer 136 (2015) E359–E386.

[2] M.F. Kaminski, M. Bretthauer, A.G. Zauber, E.J. Kuipers, H.O. Adami, M. van Ballegooijen, J. Regula, M. van Leerdam, T. Stefansson, L. Påhlman, E. Dekker, M.A. Hernán, K. Garborg, G. Hoff, I.C.C.S.G. for the Nord, The NordICC study: rationale and design of a randomized trial on colonoscopy screening for colorectal cancer, Endoscopy 44 (2012) 695–702.

[3] D. Heresbach, T. Barrioz, M. Lapalus, D. Coumaros, P. Bauret, P. Potier, D. Sautereau, C. Boustière, J. Grimaud, C. Barthélémy, Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies, Endoscopy 40 (2008) 284–290.

[4] A. Leufkens, M. van Oijen, F. Vleggaar, P. Siersema, Factors influencing the miss rate of polyps in a back-to-back colonoscopy study, Endoscopy 44 (2012) 470–475.

[5] Y. Mori, S.-e. Kudo, T.M. Berzin, M. Misawa, K. Takeda, Computer-aided diagnosis for colonoscopy, Endoscopy 49 (2017) 813–819.

[6] S.A. Karkanis, D.K. Iakovidis, D.E. Maroulis, D.A. Karras, M. Tzivras, Computer-aided tumor detection in endoscopic video using color wavelet features, IEEE Trans. Inf. Technol. Biomed. 7 (2003) 141–152.

[7] R. Nawarathna, J. Oh, J. Muthukudage, W. Tavanapong, J. Wong, P.C. De Groen, S.J. Tang, Abnormal image detection in endoscopy videos using a filter bank and local binary patterns, Neurocomputing 144 (2014) 70–91.

[8] A.V. Mamonov, I.N. Figueiredo, P.N. Figueiredo, Y.H.R. Tsai, Automated polyp detection in colon capsule endoscopy, IEEE Trans. Med. Imaging 33 (2014) 1488–1502.

[9] S.H. Bae, K.J. Yoon, Polyp detection via imbalanced learning and discriminative feature learning, IEEE Trans. Med. Imaging 34 (2015) 2379–2393.

[10] Y. Wang, W. Tavanapong, J. Wong, J. Oh, P.C. De Groen, Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy, IEEE J. Biomed. Health Inf. 18 (2014) 1379–1389.

[11] J. Bernal, J. Sanchez, F. Vilarino, Towards automatic polyp detection with a polyp appearance model, Pattern Recognit. 45 (2012) 3166–3182.

[12] N. Tajbakhsh, S.R. Gurudu, J.M. Liang, Automated polyp detection in colonoscopy videos using shape and context information, IEEE Trans. Med. Imaging 35 (2016) 630–644.

[13] M. Hafner, R. Kwitt, A. Uhl, F. Wrba, A. Gangl, A. Vecsei, Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps, Pattern Recognit. 42 (2009) 1180–1191.

[14] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[15] H.R. Roth, L. Lu, J.M. Liu, J.H. Yao, A. Seff, K. Cherry, L. Kim, R.M. Summers, Improving computer-aided detection using convolutional neural networks and random view aggregation, IEEE Trans. Med. Imaging 35 (2016) 1170–1181.

[16] N. Tajbakhsh, K. Suzuki, Comparing two classes of end-to-end machine-learning models in lung nodule detection and classification: MTANNs vs. CNNs, Pattern Recognit. 63 (2017) 476–486.

[17] A. Esteva, B. Kuprel, R.A. Novoa, J. Ko, S.M. Swetter, H.M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118.

[18] R. Zhang, Y. Zheng, T.W.C. Mak, R. Yu, S.H. Wong, J.Y.W. Lau, C.C.Y. Poon, Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain, IEEE J. Biomed. Health Inf. 21 (2017) 41–47.

[19] Z.C. Jiao, X.B. Gao, Y. Wang, J. Li, A parasitic metric learning net for breast mass classification based on mammography, Pattern Recognit. 75 (2018) 292–301.

[20] M.J. Afridi, A. Ross, E.M. Shapiro, On automated source selection for transfer learning in convolutional neural networks, Pattern Recognit. 73 (2018) 65–75.

[21] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, J. Liang, Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Trans. Med. Imaging 35 (2016) 1299–1312.

[22] Z. Yuan, M. IzadyYazdanabadi, D. Mokkapati, R. Panvalkar, J.Y. Shin, N. Tajbakhsh, S. Gurudu, J. Liang, Automatic polyp detection in colonoscopy videos, Medical Imaging 2017: Image Processing, Spie-Int Soc Optical Engineering, 2017.

[23] N. Tajbakhsh, S.R. Gurudu, J. Liang, Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks, in: 2015 IEEE 12th International Symposium on Biomedical Imaging, New York, NY, USA, 2015, pp. 79–83.

[24] S.Y. Park, D. Sargent, Colonoscopic polyp detection using convolutional neural networks, Medical Imaging 2016: Computer-Aided Diagnosis, Spie-Int Soc Optical Engineering, 2016.

[25] S. Park, M. Lee, N. Kwak, Polyp Detection in Colonoscopy Videos Using Deeply-Learned Hierarchical Features, Seoul National University, 2015.

[26] L. Yu, H. Chen, Q. Dou, J. Qin, P.A. Heng, Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos, IEEE J. Biomed. Health Inf. 21 (2017) 65–75.

[27] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2016, pp. 779–788.

[28] M. Danelljan, G. Bhat, F.S. Khan, M. Felsberg, ECO: efficient convolution operators for tracking, in: 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 6931–6939.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2016, pp. 770–778.

[30] J. Bernal, N. Tajkbaksh, F.J. Sánchez, B.J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, Comparative validation of polyp detection methods in video colonoscopy: results from the MICCAI 2015 endoscopic vision challenge, IEEE Trans. Med. Imaging 36 (2017) 1231–1249.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (2015) 211–252.

[32] D. Colussi, G. Brandi, F. Bazzoli, L. Ricciardiello, Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention, Int. J. Mol. Sci. 14 (2013) 16365–16385.

[33] F. Bishehsari, M. Mahdavinia, M. Vacca, R. Malekzadeh, R. Mariani-Costantini, Epidemiological transition of colorectal cancer in developing countries: Environmental factors, molecular pathways, and opportunities for prevention, World J. Gastroenterol. 20 (2014) 6055–6072.

[34] S. Ogino, A.T. Chan, C.S. Fuchs, E. Giovannucci, Molecular pathological epidemiology of colorectal neoplasia: an emerging transdisciplinary and interdisciplinary field, Gut 60 (2011) 397–411.

[35] S. Ogino, R. Nishihara, T.J. VanderWeele, M. Wang, A. Nishi, P. Lochhead, Z.R. Qian, X. Zhang, K. Wu, H. Nan, K. Yoshida, D.A. Milner, A.T. Chan, A.E. Field, C.A. Camargo, M.A. Williams, E.L. Giovannucci, The role of molecular pathological epidemiology in the study of neoplastic and non-neoplastic diseases in the era of precision medicine, Epidemiology 27 (2016) 602–611.

[36] J.C. van Rijn, J.B. Reitsma, J. Stoker, P.M. Bossuyt, S.J. van Deventer, E. Dekker, Polyp miss rate determined by tandem colonoscopy: a systematic review, Am. J. Gastroenterol. 101 (2006) 343–350.

[37] D. Shen, G. Wu, H.-I. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.

[38] B.H.K. Leung, C.C.Y. Poon, R. Zhang, Y. Zheng, C.K.W. Chan, P.W.Y. Chiu, J.Y.W. Lau, J.J.Y. Sung, A therapeutic wireless capsule for treatment of gastrointestinal haemorrhage by balloon tamponade effect, IEEE Trans. Biomed. Eng. 64 (2017) 1106–1114.

Ruikai Zhang received the B.E degree in Biomedical Engineering from Shenzhen University, China in 2012, and M.E degree in Biomedical Engineering from Stony Brook University, New York, USA in 2014. He is currently a PhD candidate in the Department of Surgery, The Chinese University of Hong Kong. He was a visiting student to The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA in 2017. His research interests include wearable and ingestible devices, endoscopic and health informatics.

Yali Zheng received the B.E. degree in Electronic Science and Technology from Beijing Jiaotong University in 2007, M. Phil. degree in Microelectronics and Solid States Electronics from Peking University in 2010, and Ph.D degree in Electronic Engineering from The Chinese University of Hong Kong (CUHK) in 2014. She is currently a postdoctoral fellow at Department of Surgery of CUHK.

She serves as a member of IEEE EMBS Technical Committee on Wearable Biomedical Sensors and Systems and also reviewer for a number of prestigious international journals, including the IEEE Transactions on Biomedical Engineering and IEEE Journal of Biomedical and Health Informatics. She has published over 10 SCI papers and received an h-index of 9. One of her first-author papers on “Unobtrusive Sensing and Wearable Devices for Health Informatics” has received enough citations to place it in the top 1% of the academic field of Engineering in ISI Web of Science during 2016–2018. Her current research interests include unobtrusive sensing, wearable and ingestible medical devices, and imaging informatics.

Carmen CY Poon graduated from the Engineering Science (Biomedical) Program and obtained her master degree at University of Toronto. She completed her Ph.D. in Electronic Engineering at The Chinese University of Hong Kong, where she co-founded the Division of Biomedical Engineering Research at the Department of Surgery. She has published > 100 Scopus-indexed articles and received an h-index of 20. Her works have been collectively cited > 1920 times. Her research interests include wearable sensing and endoscopic informatics that have potential to change clinical practices.

Carmen has served as IEEE EMBS AdCom (2014–2016), and Chair for EMBS Technical Committee on Wearable Biomedical Sensors and Systems (2016–2017). She is currently a member of the IEEE Biomedical Engineering Award Committee (2018–2019) and serves as an Editorial Board Member for three international journals. She was the first person to receive Early Career Award from two world’s largest international professional societies of biomedical engineers, the International Federation of Medical and Biological Engineering / The International Academy of Medical and Biological Engineering (IFMBE/IAMBE) and the IEEE Engineering in Medicine and Biology Society (EMBS).

Dinggang Shen is currently a Jeffrey Houpt Distinguished Investigator, and a Professor of Radiology, Computer Science, and Biomedical Engineering, with the Biomedical Research Imaging Center (BRIC), The University of North Carolina at Chapel Hill. He is currently directing the Image Display, Enhancement, and Analysis Lab, Center for Image Analysis and Informatics, Department of Radiology, and also the medical image analysis core at the BRIC. He was a tenure-track Assistant Professor with the University of Pennsylvanian, Philadelphia, PA, USA, and a Faculty Member with the Johns Hopkins University, Baltimore, MD, USA. He has authored over 800 papers in the international journals and conference proceedings. His research interests include medical image analysis, computer vision, and pattern recognition. He is a fellow of The American Institute for Medical and Biological Engineering, a fellow of The International Association for Pattern Recognition, and also a fellow of IEEE. He serves as an Editorial Board Member for eight international journals. He has also served on the Board of Directors, The Medical Image Computing and Computer Assisted Intervention Society, from 2012 to 2015.

James YW Lau is currently the Chairman and Yao Ling Sun Professor of Surgery to the Department of Surgery, The Chinese University of Hong Kong. He is also the Director to Endoscopy Centre, Prince of Wales Hospital.

Professor Lau graduated with honors from the University of New South Wales, Australia in 1987. He became a fellow to the Royal College of Surgeons, Edinburgh in 1991. Professor Lau then worked with Professor Sydney Chung at the Prince of Wales Hospital, where he received much of his training in upper gastrointestinal surgery and therapeutic endoscopy. His main research focuses on the multidisciplinary treatment of bleeding peptic ulcer. His research led to serval publications in international journals and a Doctoral degree in Medicine.