Improving Automatic Polyp Detection Using CNN by Exploiting Temporal Dependency in Colonoscopy Video

Hemin Ali Qadir, Ilangko Balasingham, *Senior Member, IEEE,* Johannes Solhusvik, *Senior Member, IEEE,* Jacob Bergsland, Lars Aabakken, and Younghak Shin, *Member, IEEE*

***Abstract*—Automatic polyp detection has been shown to be difficult due to various polyp-like structures in the colon and high interclass variations in polyp size, color, shape, and texture. An efficient method should not only have a high correct detection rate (high sensitivity) but also a low false detection rate (high precision and specificity). The state-of- the-art detection methods include convolutional neural net- works (CNN). However, CNNs have shown to be vulnerable to small perturbations and noise; they sometimes miss the same polyp appearing in neighboring frames and produce a high number of false positives. We aim to tackle this prob- lem and improve the overall performance of the CNN-based object detectors for polyp detection in colonoscopy videos. Our method consists of two stages: a region of interest (RoI) proposal by CNN-based object detector networks and a false positive (FP) reduction unit. The FP reduction unit exploits the temporal dependencies among image frames in video by integrating the bidirectional temporal informa- tion obtained by RoIs in a set of consecutive frames. This information is used to make the final decision. The exper- imental results show that the bidirectional temporal infor- mation has been helpful in estimating polyp positions and accurately predict the FPs. This provides an overall perfor- mance improvement in terms of sensitivity, precision, and specificity compared to conventional false positive learn- ing method, and thus achieves the state-of-the-art results on the CVC-ClinicVideoDB video data set.**

***Index Terms*—Colonoscopy, polyp detection, computer aided diagnosis, convolutional neural networks, false posi- tive learning, transfer learning, temporal information.**

I. Introduction

COLORECTAL cancer (CRC) is the second leading cause of cancer-related death in the USA for both genders, and its incidence increases, with 140,250 new cases and 50,630 deaths expected by 2018 [1]. Most colorectal cancers are adenocarcinomas developing from adenomatous polyps. Although adenomatous polyps are initially benign, they might become malignant over time if left untreated [2]. Colonoscopy is a widely used technique for screening and preventing polyps from becoming cancerous [3]. However, it is dependent on highly skilled endoscopists, and recent clinical studies have shown that 22%–28% of polyps are missed in patients undergoing colonoscopy [4]. A missed polyp can lead to late diagnosis of colon cancer and survival rates become as low as 10% [5].

Over several decades, methods based on computer vision and machine learning have been proposed for automatic detection of polyps [6]–[23]. In early studies, hand-craft features, such as color wavelet, texture, Haar, histogram of oriented gradients (HoG) and local binary pattern (LBP) were investigated [6]–[11]. More sophisticated algorithms were proposed in [12] and [13]; where valley information based on polyp appearance was used in the former and edge shape and context information were used in the later. These feature patterns are frequently similar in polyp and polyp-like normal structures, resulting in decreased performance.

Convolutional neural networks (CNN) lead to promising results in polyp detection [14]–[21]. In the MICCAI 2015 polyp detection challenge, CNN features outperformed hand-craft features [14]. However, several recent studies demonstrated that deep neural networks (DNN) including CNNs are highly vulnerable to perturbations and noise [24]–[29]. Jiawei Su *et al.* [29] have shown that current DNNs are even vulnerable to small attacks and can easily be fooled just by adding relatively small perturbations (one pixel) to the input image. Because of this

This work was supported by Research Council of Norway through the industrial Ph.D. project under the contract num- ber 271542/O30 and through the MELODY project under the contract number 225885/O70. *(Corresponding author: Younghak Shin.)*

H. A. Qadir is with OmniVision Technologies, the Intervention Cen- tre, Oslo University Hospital, 0337 Oslo, Norway, and also with the De- partment of Informatics, University of Oslo, 0315 Oslo, Norway (e-mail: hemina.qadir@gmail.com).

I. Balasingham is with Intervention Centre, Oslo University Hospital 0337, Oslo Norway, and also with the Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Nor- way (e-mail: ilangko.balasingham@medisin.uio.no).

J. Solhusvik is with OmniVision Technologies, 0337 Oslo, Norway, and the Department of Informatics, University of Oslo, 0315 Oslo, Nor- way (e-mail: johannes.solhusvik@ovt.com).

J. Bergsland is with Intervention Centre, Oslo University Hospital, 0337 Oslo, Norway (e-mail: jacobbergsland622@gmail.com).

L. Aabakken is with the Department of Transplantation, Faculty of Medicine, University of Oslo, 0315 Oslo, Norway, and Oslo University Hospital, 0337 Oslo, Norway (e-mail: larsaa@medisin.uio.no).

Y. Shin is with the Department of Electronic Systems, Norwegian University of Science and Technology, 7491 Trondheim, Norway (e-mail: shinyh0919@gmail.com).

vulnerability, CNN networks might be fooled by the specular highlights and small changes in polyp (other elements) structures appearance in colonoscopy. This means the CNN networks can easily miss the same polyp appearing in a sequence of neighboring frames and produce unstable detection output contaminated with a high number of FPs. To the best of our knowledge, this paper is the first to study the CNN' s vulnerability in polyp detection.

In this paper, we aim to tackle these problems by exploiting the temporal dependencies among consecutive frames. We propose a method to find and remove FPs and detect intra-frame missed polyps based on the consecutive detection outputs of CNN-based detectors. The hypothesis is that neighboring frames should contain the same polyp, and the detected polyp should be closely similar in position and size. We use a dataset of still images for training, and make the trained models useful for polyp detection in colonoscopy video. At inference time, we can take advantage of the multitude of detected bounding boxes in consecutive frames. We use bidirectional temporal coherence information from the detection outputs to make the final decision for the current frame. This approach can improve the sensitivity, precision, and specificity of the detector models. We can also stabilize the detection outputs by forcing the system to find the missed polyps and refine the detection coordinates within a sequence of frames. We demonstrate that the proposed method outperforms the results obtained with state-of-the-art object detectors, i.e., faster region based convolutional neural network (Faster R-CNN) [30] and single shot multibox detector (SSD) [31].

II. Related Work

From a clinical perspective, performance of a given computer-aided diagnostic tool should have high sensitivity (high true positive rate, TPR) and high precision (low false positive rate, FPR) [23]. Low sensitivity is unacceptable since it gives a false sense of security while low precision affects the psyche of the patients and annoys clinicians. In the large bowel, there are various structures of normal mucosa that closely resemble the characteristics of polyps. This makes polyp detection task more difficult for both CNN and hand-craft features, resulting in the present low precision rates.

Recently, Dou *et al.* [32] proposed false positive learning (FP learning) to reduce FPs and increase precision in Cerebral Microbleeds detection from MR images. Shin *et al.* [15] and Angermann *et al.* [22] adapted FP learning for polyp detection. Although FP models can successfully decrease FPs, true positive (TP) detections decline [15], [22]. In this work, we propose an efficient FP reduction method which improve both sensitivity and precision. Later, we also validate our method on FP models for further performance improvement.

Another active method to reduce FPs is to include time information during detection in video sequences [11], [16]–[18], [23]. Sun *et al.* [11] used the previous and the future frames to model the probabilistic dependence between adjacent frames using conditional random fields with the Markov property. Angermann *et al.* [23] extended their previous work [22] by adding a spatio-temporal module to incorporate temporal coherence information from the two previous frames. Tajbakhsh *et al.* [16] and Zhang *et al.* [17] incorporated information from the detection in the previous frames to enhance the polyp detection performance. In [17], an online object tracker was used in combination with YOLO [33] to increase sensitivity, more TPs. This model failed to increase both precision and specificity due to the introduction of new FPs. The main reason for these new FPs could be the lack of temporal information fed into the tracker as it relies on previous frames only. When FPs are used to initialize the tracker more FPs will be generated. Yu *et al.* [18] proposed a 3D fully convolutional network (FCN) framework to learn spatio-temporal features from volumetric data and generate more discriminative features [34]. They extracted a video clip of 16 frames (7 previous and 8 future frames) to train an offline and online 3D-FCNs. This method is computationally expensive and needs 1.23 sec (beside the delay from using future frames) to generate the final decision. Unlike [17] and [18], we use 3D temporal information extracted from a video clip after a 2D-CNN is applied to provide RoIs for each frame. We use temporal dependencies among future and previous frames to more reliably filter out FPs and Keep TPs.

III. Methods

The proposed system consists of two stages: a RoI proposal network stage, and FP reduction stage (see Fig. 1). In the first stage, a CNN based detector, e.g., Faster R-CNN and SSD, suggests multiple RoIs to the next stage. In the second stage, the proposed RoIs of the current frames are examined and categorized as TPs or FPs by considering the RoIs of some consecutive frames.

A. *The RoI Proposal Network*

The RoI Proposal Network is a CNN-based detector model able to propose a number of RoIs for the FP reduction unit. For each frame, the detector can generate up to 100 RoIs and sort them based on their confidence values in which the top one has the highest value. At test time, we control how many RoIs are considered for the next stage. There is a trade off between sensitivity and precision relative to the number of RoIs considered, i.e., a large number of RoIs causes higher sensitivity but lower precision.

The RoI proposal network can be any CNN-based detector model. In this study, we only consider Faster R-CNN [30] and SSD [31] architectures to investigate polyp detection performance improvement using our method. In fact, these two detector models can be utilized as a standalone model for automatic polyp detection. Both detector architectures are designed for object detection in a single independent frame, and have no mechanism to adapt temporal information during training and testing phases. They produce a high number of FPs and may miss the same polyp appearing in neighboring frames. In Section V, we will show the results of these detectors when used alone and compare them to the results obtained with our proposed method.

In these detector models, a collection of boxes acting as anchors are overlaid on the image at different spatial locations,

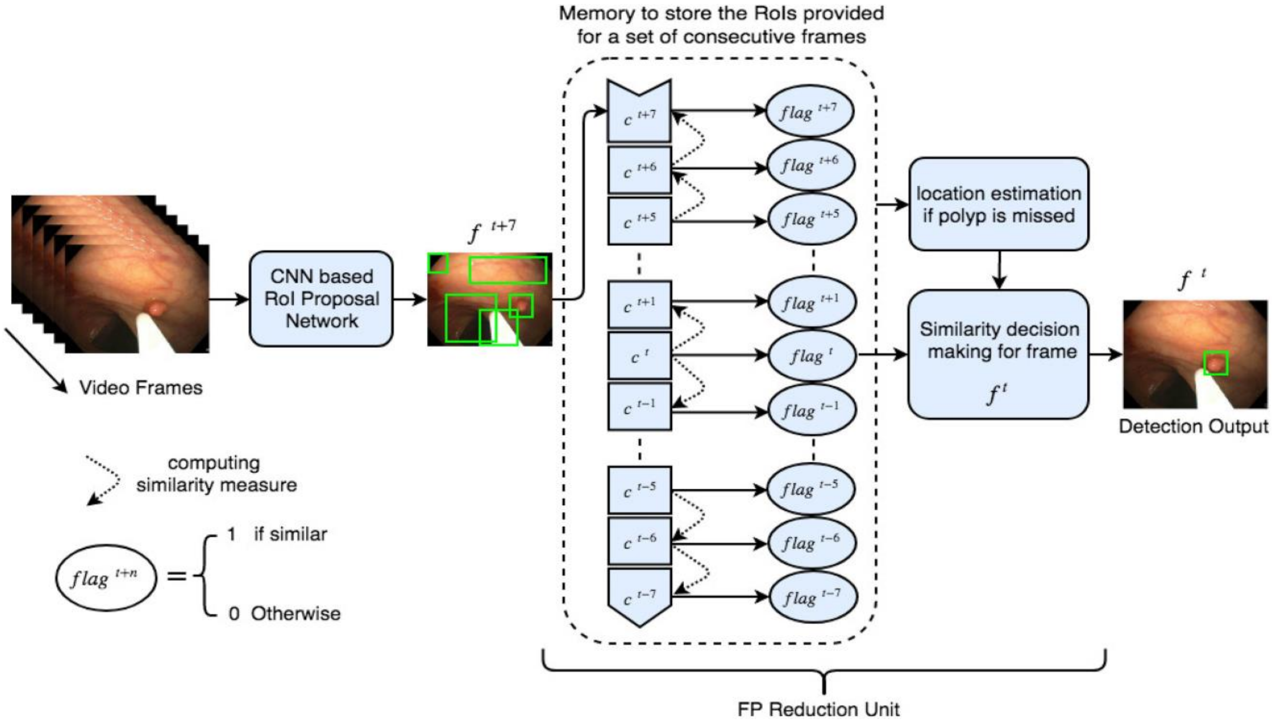


Fig. 1. Procedure of the proposed system. The CNN-based proposal network provides RoIs to the FP reduction unit. The FP reduction unit performs the following: 1) classifies the proposed RoIs as TPs or FPs using a similarity measure to find temporal coherence among a set of consecutive frames, 2) estimates the location of missed polyps using interpolation.

scales, and aspect ratios [30], [31]. Then, a model is trained to predict: category scores for each anchor, and a continuous box offset by which the anchor needs to be shifted to fit the ground-truth bounding box. The objective loss function is a combined loss of classification and regression losses. For each anchor a, the best matching ground-truth box b will be found. If there is such a match, anchor a acts as a positive anchor, and we assign a class label $y_a \in \{1, 2, ...K\}$, and a vector $(\varphi(b_a; a))$ encoding box b with respect to anchor a. If there is no match, anchor a acts as a negative sample, and the class label is set to $y_a = 0$. The loss for each anchor a, then consists of two losses: location-based loss f_{loc} for the predicted box $f_{loc}(I; a, \theta)$, classification loss f_{cls} for the predicted class $f_{cls}(I; a, \theta)$, where I is the image and θ is the model parameter, the overall loss function to train a model is to minimize a weighted sum of the localization loss and the classification loss over a mini-batch of size m

$$\mathrm{L}(a,I;\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N}\sum_{j=1}^{N}\alpha\cdot 1[a\,is\,positive].f_{loc}\left(\varphi(b_a;a)-f_{loc}(I;a,\theta)\right)+\beta\cdot f_{cls}\left(y_a,f_{cls}(I;a,\theta)\right), \quad (1)$$

where N is the number of anchors for each frame, and α, β are weights balancing the localization and the classification loss. For both models, we use the Smooth L1 loss [35] for computing the localization loss between the predicted box and the ground-truth box. The classification loss is the softmax loss.

1) Faster R-CNN: To detect objects in an image, Faster R-CNN uses two stages: region proposal network (RPN), and a box classifier network. Both networks share a common set of convolutional layers to reduce the marginal cost for computing region proposals. The RPN utilizes feature maps at one of the intermediate layers (usually the last convolutional layer) of the CNN feature extractor network to generate class-agnostic box proposals, each with an objectness confidence value. The proposed boxes are a grid of anchors titled in different aspect ratios and scales. The box classifier network uses these anchors to crop features from the same intermediate feature map and feeds the cropped features to the remainder of the network in order to predict object categories and offsets in bounding box locations. The loss functions for both stages take the form of Eq. (1).

The RPN can benefit from deeper and more expressive features because it learns to propose regions from the training data [30]. By using Faster R-CNN, we aim to design a highly accurate polyp detector and show that its results can be improved with the proposed method. We decide to use a very deep network—Inception Resent [36]—as the feature extractor network. The RPN generates 300 proposals from the "Mixed_6a" layer including its associated residual layers. Unlike [30], we use "crop_and_resize" operation in Tensorflow instead of RoI pooling [37]. During training, the anchors are classified as either negative or positive samples based on Jaccard overlap matching. Shin *et al.* [15] evaluated different Jaccord overlap thresholds for polyp detection and recommended 0.3 and 0.6 to choose negative and positive samples respectively. After the matching step, most of the anchors are negatives. Instead of using all the negative samples, we set the ratio between negatives and positives to 1:1 to avoid imbalance training. In Faster R-CNN, models are trained on image resized to M on the shorter edge. For our polyp model, we set M to be the height of the training images to keep the original image size.

2) SSD: Unlike Faster R-CNN, The SSD approach uses a single deep neural network for object detection in an image and eliminates the need for an extra proposal generation network. This makes SSD a much faster object detector than Faster R-CNN. To handle objects of various sizes and achieve higher detection accuracy, SSD evaluates a fixed set of anchor boxes of different aspect ratios at multiple feature maps from multiple layers to predict the category scores and box offset. In SSD, the input images are always re-sized to $M \times M$ pixel resolutions. Image resolution is a way to trade accuracy for speed—higher resolution means higher accuracy, but lower detection speed. We set $M = 600$ for our SSD model. The purpose of using SSD in our study is to show that the proposed method is effective for less accurate object detector. We choose MobileNet [38] as the CNN feature extractor, and follow the methodology in [31] to generate anchors by selecting the topmost convolutional feature maps (*conv*–1 and *conv*–3) and appending four additional convolutional layers with spatial decaying resolution with depths 512, 256, 256, 128 respectively. We use ReLU6 in all layers except the softmax layer. During training, we treat those anchors with Jaccard overlap higher than a threshold of 0.5 as positive anchors and the rest as negatives. We set the ratio between negatives and positives to 3:1, recommended ratio by the original paper [31].

B. FP Reduction Unit

In the FP reduction unit, we identify detection irregularities and outliers in a video sequence. When a polyp appears in a sequence of frames, its location slightly changes following a motion estimating the movement in the sequence. Irregularities and outliers are those detection outputs that do not smoothly follow such a movement. More specifically, outliers are those outputs that appear to be FPs among a set of TPs (see Fig. 3b). The proposed RoIs in a number of consecutive frames are passed through another process to find irregular detection outputs before the final decision is made for the RoIs in the current frame. We consider those detection irregularities and outliers as FPs. In case of an outlier, an action is taken to correct the detection. Therefore, the FP reduction unit comprises of two processes: a mechanism to detect FPs, and a mechanism to correct the outliers denoting the missed polyps in the sequence.

1) FP Detection Mechanism: To detect irregularities and outliers, we use the coordinates provided by the RoI proposal network as features. Fig. 2 presents the coordinate points of a proposed RoI used in this study to collect 8 features—$x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$, x_c, y_c, w, and h. We use all these coordinate points to detect even small irregularities in the detection outputs and refine them if they appear to be outliers (see Fig. 12a and Fig. 12c). To handle different frame sizes, we normalize the coordinate points by dividing them by the frame width and height.

A distance metric (e.g., Euclidean distance) can be applied to compute the similarity measure between the features of RoIs provided for a set of consecutive frames. Only those RoIs with high similarity measure (smaller than a distance threshold value) should be considered to generate the final detection output in the current frame, and those RoIs without spatio-temporal overlap (higher than the distance threshold value) should be eliminated for the final decision.

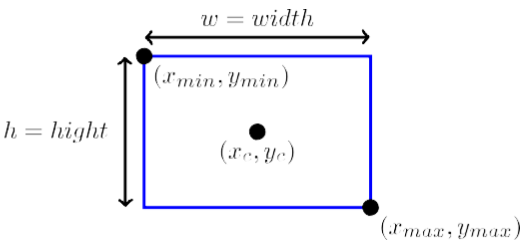


Fig. 2. Coordinates of a RoI used as features.

We propose an algorithm shown in Fig. 1 in which some previous and future frames are considered in order to choose the proposed RoIs as true detection outputs in the current frame—the frame in the middle. The question regarding how many frames need to be considered is an optimization problem that we will discuss later in Section IV-E. The optimal number is 15 (see Fig. 5) consecutive frames i.e., 7 previous frames and 7 future frames. The CNN-based detector in the first stage continuously generates RoIs for the last frame. We store the features of each RoI of the 15 consecutive frames in a matrix called c. The size of matrix c depends on the number of RoIs (r) provided per frame and the number of frames considered (f). The matrix size is $f \times r \times d$ where d is the dimension of the features, 8 in our case.

For the sake of simplicity, we only show the contents of matrix c when one RoI per frame is provided. This will allow us to write the mathematical equations in simpler forms. Matrix c for one RoI per frame can then be expressed as follows

$$\left.\begin{aligned} &c = [c^{t-7} \ldots\ldots c^{t-2} c^{t-1} \mathbf{c^t} c^{t+1} c^{t+2} \ldots\ldots c^{t+7}]^T, \\ &c^{t+n} = [x^{t+n}_{\min}\ y^{t+n}_{\min}\ x^{t+n}_{\max}\ y^{t+n}_{\max}\ x_c^{t+n}\ y_c^{t+n}\ w^{t+n}\ h^{t+n}], \\ &n \in \{-7, -6, \ldots\ldots, -2, -1, 0, 1, 2, \ldots\ldots, 6, 7\}. \end{aligned}\right\} \tag{2}$$

At an initial study, we used the Euclidean distance as the similarity metric, later we optimize the proposed model by evaluating several distance metrics. Using the Euclidean distance, the similarity between two RoIs of two consecutive frames (f^t and f^{t+1}) is measured as follows

$$d_2 : (c^t, c^{t+1}) \mapsto \|c^t - c^{t+1}\|_2 = \sqrt{\sum_i (c_i^t - c_i^{t+1})^2}, \quad c_i \in \{x_{\min}, y_{\min}, x_{\max}, y_{\max}, x_c, y_c, h, w\}. \tag{3}$$

Every time, the RoIs provided for the current frame f^t are compared to the RoIs in the previous frame f^{t-1} and the future frame f^{t+1}. If the similarity measure for a particular RoI in either

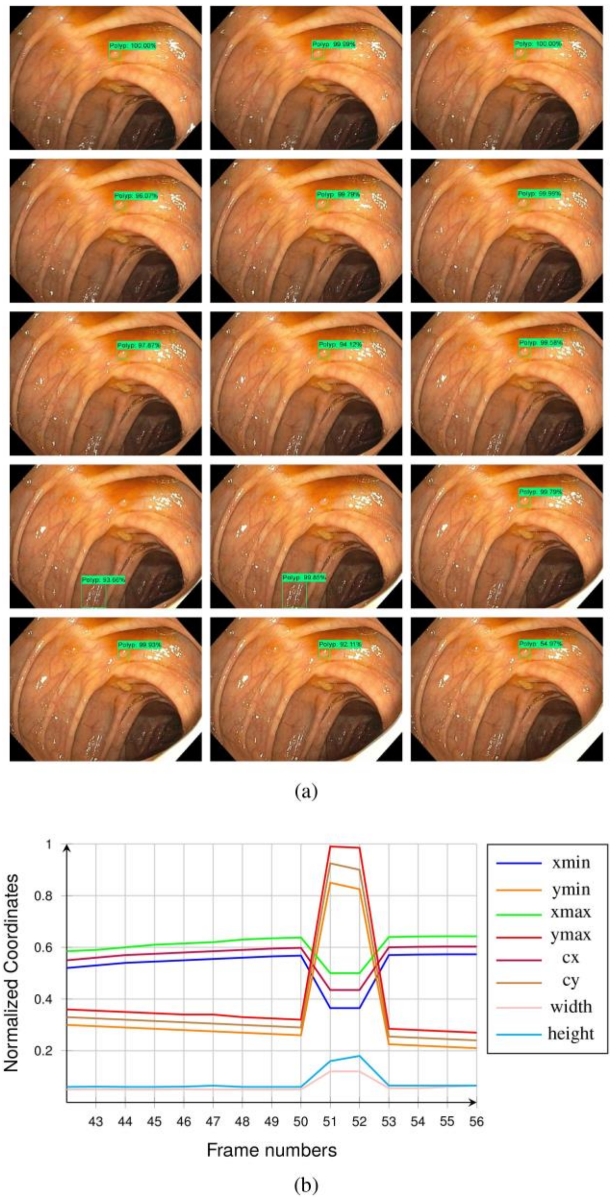


Fig. 3. A sequence of frames starting from frame 42 (top left frame) and ending at 56 (bottom right frame) shows a case where the same polyp is missed in frames 51 and 52. (a) detection results in the sequence (b) the normalized coordinates of the proposed RoIs in the sequence. In (b), The coordinates of frame 51 and 52 are two outliers compared to the other detected RoIs, and thus can be considered as FPs.

direction is smaller than a threshold value, the flag corresponding to that RoI is set to 1. Otherwise, the corresponding flag is set to 0. The number of flags for each frame is equal to the number of RoIs provided by the CNN detector, therefore, the size of the *flags* matrix is $f \times r$. The other frames in the set only need to be checked with one frame in one direction. For instance, frame f^{t+1} needs to be checked with frame f^{t+2} , and the corresponding flags are set based on the similarity measure. If no similar RoI found in frame f^{t+2} , frame f^{t+1} will be checked with frame f^{t+3} , and all the corresponding flags for frame f^{t+2} will be set to 0. This checking process continues until the last two frames in both directions are reached.

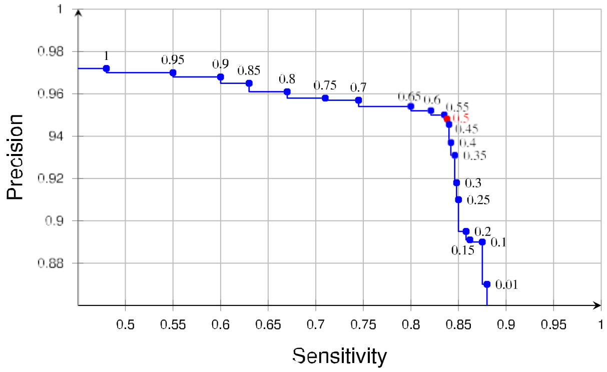


Fig. 4. Truncated Precision-Sensitivity curve showing the effect of changing *avg–th* on the performance. The numbers shown above the curve are the *avg–th* values. 0.5 is chosen to keep the balance between precision and sensitivity.

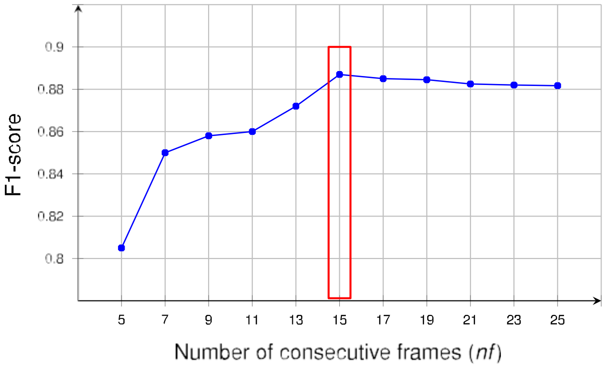


Fig. 5. F1-score when the number of frames (*nf*) is varied. F1-score is maximum when *nf* is 15 frames.

Once all the flags are set, we classify each RoI provided for the current frame. If the number of flags with value 1 accumulated for a specific RoI is less than 7, this RoI is classified as FP, and thus it will be deleted. In other words, we only pick those RoIs overlapped with at least 7 RoIs in a set of 15 consecutive frames. Furthermore, we calculate the average confidence for the overlapped RoIs and only classify those RoIs with an average confidence (*avg–th*) ≥ 0.5 (an optimized value, see Fig. 4) as TPs. In this way, we have less FPs and keep only those RoIs that repeat in more than 7 consecutive frames with high confidence values in the final output.

2) Correction Mechanism: Since the CNN detectors are vulnerable to small variation, the same polyp might be missed in a couple of frames in a video sequence. Fig. 3a presents a case where the same polyp is correctly detected by the CNN-based detector in most of the frames but missed in a couple of frames in the sequence (i.e., frame 51 and 52). In Fig. 3b, we can clearly see these outliers in the curves drawn from the eight coordinate points of the provided RoIs.

When outliers are detected, the correction mechanism can be performed on future frames before they become the current frame in the sequence. In particular, we only apply the correction mechanism when the missing occurs in frames f^{t+1} ,f^{t+2} ,f^{t+3} , or/and f^{t+4} . The other two important conditions to apply the correction mechanism are: the number of flags with value 1 accumulated during the FP detection process for a specific RoI has to be larger than 7 (optimized number), and at least there is a RoI in the next frames coincident with RoIs in the previous frames. If all these conditions are met, we set the outlier data points to zeros in matrix c based on the flag sets. That means we will have missing points in the data points representing the coordinates of the RoIs in matrix c. Now we have a function that is only known at a discrete set of data points (f^{t+n}, c^{t+n}). We can use interpolation to estimate the values of that function at frames of f^{t+n} not included in the data. An interpolation function $I(f^{t+n})$ passes through the points of a discrete data set

$$I(f^{t+n}) = c^{t+n}. \tag{4}$$

Usually, we prefer a function that smoothly connects the data points. One possibility is to use the polynomial of the least degree that passes through all of the points. To find the missed polyps within inter-frames, we compute interpolation for each column in matrix c as a function of the frame number separately from each other using the Lagrange interpolation formula [39] as follows

$$I(f) = \sum_{n} c^{t+n} \prod_{n(j/=n)} \frac{f-f^{t+j}}{f^{t+n}-f^{t+j}}. \tag{5}$$

This results in a continuous and smoothed curve. This function can simply estimate the polyp position in the sequence, mainly due to the use of the future frames to estimate the location of missed polyps in inter-frames in the sequence. The confidence values for the new generated RoIs are also calculated using Eq. (5). We illustrate the proposed method in pseudocode shown in Algorithm 1 to summarize and describe the entire procedure.

IV. Experimental Setup

A. Experimental Datasets

We used three publicly available datasets, one still frame dataset, CVC-CLINIC [12] and two colonoscopy video datasets, ASU-Mayo Clinic [13] and CVC-ClinicVideoDB dataset [23]. We used each dataset for different purposes i.e., training, validation and testing. In this way, the system will more likely be generalized because there is no any similar frames in the training and testing datasets.

CVC-CLINIC was used for training the CNN detectors, i.e., Faster R-CNN and SSD. This dataset consists of 612 Standard Definition (SD) frames of 576 × 768 pixel resolutions. The frames are extracted from 31 different videos, each containing at least a unique polyp.

ASU-Mayo Clinic is a set of 38 different and fully annotated videos. 20 videos are assigned for the training stage whereas 18 videos for testing. The ground-truth of the 18 testing videos is not publicly available. Therefore, we only used the 20 training videos, in which 10 videos are positive (with polyps) and the other 10 videos are negative (without polyps). We split the 20 training videos into validation, training and test sets. We used the 10 positive videos for validating and tuning the hyper-parameters of the proposed method. By validating and tuning the system, we aimed to find the best hyper-parameters for both the RoI proposal networks and the FP reduction unit, and realize a generalized model for other unseen datasets. We used 5 negative videos to evaluate and compare the specificity of our model and the existing FP model [15]. The remaining 5 negative videos were used for FP sample selection for the FP model.

We used CVC-ClinicVideoDB dataset to evaluate the overall performance of the proposed model. This dataset comprises of 18 videos, each with a unique polyp that appears multiple times in the videos. The total number of frames in this dataset is 11954 frames whereas only 10025 frames are annotated as having polyps. The size of the frames is 768 × 576. This dataset aims to cover all different possible scenarios that a given support system should face, making it very useful for the overall system evaluation [23].

The ground-truth for all polyp frames in all three datasets is provided. All annotations have been reviewed and corrected by clinical experts. The ground-truth provided for CVC-CLINIC and ASU-Mayo Clinic is exact boundaries around the polyp parts in the frames, while the ground-truth for polyps in CVC-ClinicVideoDB dataset is an approximation, i.e, an ellipse is drawn around the polyps.

B. Evaluation Metrics

We use the common evaluation metrics of object detection to evaluate the performance of our polyp detection method. The output of the models is four coordinates (x, y, w, h) of the detected rectangular bounding boxes. Therefore, we define the term "polyp detection" as the process of finding the polyp location within a given frame. Based on that, the following parameters are defined as follows:

True Positive (TP): True detection, the centroid of the detection falls within the polyp boundary. In case of having multiple true detection outputs for the same polyp, we will only count one TP.

True Negative (TN): True detection, no output detection for a frame without a polyp (negative frames).

False Positive (FP): False detection, the centroid of the detection falls outside the polyp boundary. In case of having multiple RoIs proposals, there can be more than one FP per frame.

False Negative (FN): False detection, the polyp is not detected in a frame containing a polyp.

Using these parameters, we can calculate the following metrics to precisely evaluate the performance:

Sensitivity: It is also called True Positive Rate (TPR) and Recall. It measures the proportion of actual polyps that are correctly detected

$$Sensitivity(Sen) = \frac{TP}{TP+FN} \times 100. \tag{6}$$

Precision: It measures how precise the model at correctly localizing a polyp within a frame

$$Precision(Pre) = \frac{TP}{TP+FP} \times 100. \tag{7}$$

Specificity: It is also called True Negative Rate (TNR). It measures the proportion of actual negative frames that are correctly

Algorithm 1: Algorithmic Framework Describing the Basic Steps of the Proposed System.

```
Input: video frames
initialize matrix $c \leftarrow 0$
for $f^t = 1$ to $M$ do {M: no. of frames in a video}
  if $f^{t+7} \in [1,2,3,4,5,6]$ then {wait till $f^1$ becames $f^t$}
    $c^{t+7} \leftarrow RoIProposalNetwork(f^{t+7})$
  else
    $c^{t+7} \leftarrow RoIProposalNetwork(f^{t+7})$
    initialize matrix $flag^t \leftarrow 0$
    $c^{next} \leftarrow c^t$
      $c^{previous} \leftarrow c^t$
      for $i = 1$ to 7 do
        if $||c^{next} - c^{t+i}||_2$ � 0.65 then {future frames}
        $flag^{t+i} \leftarrow 1$
          $c^{next} \leftarrow c^{t+i}$
        end if
        if $||c^{previous} - c^{t-i}||_2$ � 0.65 then {previous
        frames}
          $flag^{t-i} \leftarrow 1$
          $c^{previous} \leftarrow c^{t-i}$
        end if
      end for
      if $sum(flag) < 7$ then
        $c^t \leftarrow 0$ {$c^t$ is considered as FP}
      else
        keep $c^t$ {$c^t$ is considered as TP}
        if $flag^{t+1} = 0$ and $(flag^{t+2}, flag^{t+3}$ or
        $falg^{t+4}) /= 0$ then {Correction Mechanism}
          $I(f) = {}_n c^{t+n} \quad {}_{n(j=n)} \quad \frac{f-f^{t+j}}{f^{t+n}-f^{t+j}}$
        end if
      end if
    end if
    for $k = 0$ to 6 do {shift matrix $c$ to the left}
      $c^{t-k-1} \leftarrow c^{t-k}$
      $c^{t+k} \leftarrow c^{t+k+1}$
    end for
Output: $c^t$ (coordinates, confidence)
end for
```

classified

$$Specificity\,(Spec) = \frac{TN}{TN + FP} \times 100. \tag{8}$$

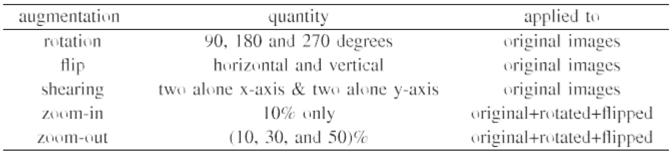
F1-score: It can be used to consider the balance between sensitivity and precision

$$F1 - score\,(F1) = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \times 100. \tag{9}$$

C. Training the Detectors

To train both CNN-based detectors, we used the CVC-CLINIC dataset. This dataset consists of 612 positive samples (images with polyps). This low number of images is not sufficient to train deep neural networks [40]. To prevent the detectors from overfitting and enlarge the training samples, we utilized different augmentation strategies. It is important to apply the augmentation strategies by considering real colonoscopy scenarios and variations that a given system will face. In real colonoscopy recordings, polyps show large inter-class variation such as changes in colors, scales, and positions in addition to changes in viewpoints due to camera movement. To cover these variations, we applied not only image rotation and flipping but also zoom-in, zoom-out, and shearing. Table I presents all the augmentation techniques applied to enlarge the training dataset.

TABLE I
AUGMENTATION STRATEGIES APPLIED TO ENLARGE THE DATASET



augmentation	quantity	applied to
rotation	90, 180 and 270 degrees	original images
flip	horizontal and vertical	original images
shearing	two alone x-axis & two alone y-axis	original images
zoom-in	10% only	original+rotated+flipped
zoom-out	(10, 30, and 50)%	original+rotated+flipped

The reason for having three zoom-out and only one zoom-in is that detection of small size polyps is more difficult compared to large size polyps. With this imbalance zooming, we can enforce the detectors to find small size polyps more efficiently. We excluded those polyps that disappeared after applying zoom-in. The total number of training samples became 18594 images after applying the augmentation methods presented in Table I.

Even though the dataset is enlarged, it does not guarantee that the proposed model is prevented from overfitting and performs well in the test phase. The main reason is that the training dataset contains only 31 different unique polyps, and augmentation methods do not improve data distribution, they only lead to an image-level transformation through depth and scale. To overcome the lack of training data in medical applications, N. Tajbakhsh et al. [40] demonstrated that pre-trained CNN feature extractors with proper fine-tuning can outperform training from scratch. We therefore used transfer learning by initializing weights of the CNN feature extractors with pre-trained models. Both CNN feature extractors were trained on Microsoft’ s COCO (Common Objects in Context) dataset [41], using all 80 K samples of “2014 train” and a subset from 32 K samples of “2014 val”, holding 8000 examples for validation [37].

We fine-tuned the pre-trained models using the augmented dataset. For Faster R-CNN, we used SGD with a momentum of 0.9 and batch sizes of 1. We set the maximum number of epochs to 30 with the learning rate equal to 0.0001. For SSD, we used RMSProp [42] with a decay of 0.9 and batch sizes of 18. Since the SSD converges slower than Faster R-CNN, we needed to take more epochs. We set the maximum number of epochs to 300 with the learning rate of 0.002.

D. False Positive Models

From a clinical perspective, high precision is desirable, but this is difficult in automatic polyp detection. There are various structures which closely resemble polyp characteristics [14], resulting in performance degradation especially in precision. Using only positive samples to train a detector model, negative samples are selected from the background during training.

To avoid imbalance training, only a portion of the background patches that have zero or small Jaccard overlap (<0.5 for SSD, and <0.3 for Faster R-CNN) with polyp masks will be considered as negative samples [30], [31]. In this way, it is difficult to have exact bounding boxes around structures mimicking polyps, and the two polyp detector models do not efficiently learn how the hard negative samples would look like [15], [22]. Therefore, they tend to generate many FPs (see the result in Section V).

For comparison, we followed the procedure proposed by Shin *et al.* in [15] to collect strong FP samples and obtain the FP models for our polyp detectors. We set the confidence threshold to 99% and applied our two trained polyp detectors separately on 5 negative videos from ASU-Mayo Clinic dataset. For Faster R-CNN model, we collected 654 images, and for SSD model, we collected 536 images. We further increased the number of negative samples by applying 5 rotations to the collected FP samples, generating 3924 FP samples for Faster R-CNN, and 3216 FP samples for SSD. We enlarged the training dataset by combining the initial training samples (18594 positive samples) with these FP samples and their augmented ones. Using the enlarged dataset, we fine-tuned both polyp detectors to strengthen their detection capability and obtained their FP models.

E. Parameter Optimization for the Proposed Model

Before testing our models, we need to find a set of optimal parameters such as the distance threshold value (*dv*), the number of consecutive frames (*nf*) and the average confidence value (*avg–th*). A selection of the most effective distance metrics for our model can be considered as an optimization problem. We evaluate 8 commonly used distance metrics (*dm*) such as Euclidean, Manhattan, Chebyshev, Minkowski, Canberra, Cosine, Correlation and Chi-square.

We define an optimization problem P as a function of the model parameters ω which is a function of *dm*, *dv*, *nf* and *avg–th*. Since we wish to improve sensitivity and precision, and keep a balance between them, we consider P to be F1-score of the system. Therefore, the goal is to maximize P on a given validation set (S_{valid}) using a grid search on a fixed set of values for each parameter

$$\omega^*(dm,dv,nf,avg\text{–}th) = \arg\max_{\omega} P\Big(dm,dv,nf, avg\text{–}th, S_{valid}\Big). \qquad (10)$$

We used 10 positive videos of ASU-Mayo Clinic as the validation dataset (S_{valid}). Each distance metric has a different domain of acceptable values. We performed small experiments over each distance metric to find its range of acceptable values and shrink the search domain. For each distance metric *dm*, we varied the distance value *dv* in increments of a small step size. Regarding how many consecutive frames *nf* should be considered, we took 11 scenarios by changing *nf* from 5 to 25 frames in increments of 2. We let the RoI proposal network give one RoI per frame, and run this optimization problem.

We obtained the Canberra metric with $dv = 0.65$, $avg\text{–}th = 0.5$ and 15 consecutive frames as the optimal values for the purposed model. In Fig. 4, we show the precision-sensitivity curve showing the effect of the changing *avg–th*. To compute the similarity measure between two RoIs from two neighboring frames (f^t and f^{t+1}), the formula for Canberra distance metric [43] can be defined as follows

$$d_{CAD} : (c^t, c^{t+1}) \ast\longrightarrow \sum_i \frac{(c_i^t - c_i^{t+1})}{|c_i^t| + |c_i^{t+1}|}. \qquad (11)$$

Fig. 5 illustrates the effect of *nf* on F1-score. We used Canberra metric with $dv = 0.65$, and only changed *nf* from 5 to 25 frames in increments of 2. F1-score is maximum when $nf = 15$ frames. 15 is a reasonable value to keep the balance between the sensitivity and precision. When *nf* is a small number, finding FPs may become difficult as the probability of FP repetition in a small number of frames is higher than a large number. On the other hand, we may lose many TPs when *nf* is large. Since the difference between the distance metrics is not significant, we do not provide in this paper the evaluation results of the distance metrics we used.

V. Experimental Results

In this section, we present the performance of the proposed method and compare it with the performance of the original detectors, i.e., without FP reduction unit. The objective of this study is to improve sensitivity and precision. Since the proposed model is designed to find FPs, it should be able to improve the specificity. To investigate the overall performance improvement, we evaluate two datasets: 18 positive videos from CVC-ClinicVideoDB to explore the improvement in the sensitivity and precision, and 5 negative videos from ASU-Mayo Clinic to explore the improvement in the specificity.

The two detector models are able to generate up to 100 proposals per frame. They sort the proposals based on their confidence values. When we let the detectors provide one proposal per frame, the top one is returned as the detection result. Due to the existence of FPs, it is not always the case that the top detection contains the polyp. The polyp might be bounded by the second or other RoI proposals. To increase the detection capability and build a multi-polyp detection model, we need to let the detectors provide more than one RoIs per frame. Although this will enhance the sensitivity, it will degrade the precision as the majority of these 100 proposals are FPs. To further validate the capability of the proposed model, we evaluate two scenarios: one proposal per frame, and multiple proposals per frame. We later apply our FP reduction method on the results obtained by the two original detectors when their confidence threshold ($score\text{–}th = 0.5$). This is to confirm that our method is still effective in exposing FPs and maintaining TPs in the output detection of these detectors.

A. One RoI per Frame

In this scenario, we let the RoI proposal network provide one RoI per frame. The confidence threshold value of the RoI proposal network must be set to 0 so that the CNN detectors always return the top RoI regardless of its confidence value. In other words, every frame will be considered as a positive frame—assuming there are no TN frames in the videos. In case

TABLE II

Results Obtained on the 18 Positive Videos From CVC-ClinicVideoDB for One RoI per Frame Scenario: in Each Sub-Table, the 1st Row Shows the Result of the Detector Models With Score Threshold of 0.5, the 2nd Row Shows Maximum Detection Capability of the Detector Models With the Score Threshold of 0, and the 3rd Row Shows the Result of the Proposed Method Applied on the 2nd row Result

A) Faster R-CNN model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8033	1648	1151	1992	80.13	82.98	82.53
Faster R-CNN	0.0	8287	3667	0	1738	**82.66**	69.32	75.04
proposed method	0.5	8171	1166	1347	1854	81.51	**87.51**	**84.4**

B) FP model of Faster R-CNN used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
FP model [15]	0.5	6985	590	1714	3040	69.68	92.21	79.38
FP model	0.0	8259	3697	0	1768	**82.35**	69.07	75.12
proposed method	0.5	7594	576	1684	2431	75.75	**92.95**	**83.47**

C) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5443	895	1629	4582	54.29	85.88	66.53
SSD	0.0	6460	5494	0	3565	**64.44**	54.04	58.78
proposed method	0.5	5894	694	1676	4131	58.79	**89.47**	**70.96**

D) FP model of SSD used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
FP model	0.5	5023	319	1817	5002	50.10	94.03	65.37
FP model	0.0	6448	5506	0	3577	**64.32**	53.94	58.68
proposed method	0.5	5729	200	1833	4296	57.15	**96.63**	**71.82**

TABLE III

Results Obtained on the 5 Negative Videos From ASU-Mayo Clinic Dataset for One RoI per Frame Scenario: In Each Sub-Table, the 1st row Shows the Result of the Detector Models With Score Threshold of 0.5, the 2nd Row shows the Results of the Detector Models by Setting the Score Threshold to 0, and the 3rd Row Shows the Result of the Proposed Method Applied on the 2nd Row Result

A) Faster R-CNN model used as the RoI proposal network

Method	score_th	FP	TN	Spec %
Faster R-CNN [15]	0.5	2192	4662	68.02
Faster R-CNN	0.0	6854	0	0
proposed method	0.5	**1079**	**5775**	**84.26**

B) FP model of Faster R-CNN used as the RoI proposal network

Method	score_th	FP	TN	Spec %
FP model [15]	0.5	73	6781	98.93
FP model	0.0	6854	0	0
proposed method	0.5	**8**	**6846**	**99.88**

C) SSD model used as the RoI proposal network

Method	score_th	FP	TN	Spec %
SSD	0.5	1096	5758	84.01
SSD	0.0	6854	0	0
proposed method	0.5	**435**	**6419**	**93.65**

D) FP model of SSD used as the RoI proposal network

Method	score_th	FP	TN	Spec %
FP model	0.5	264	6590	96.15
FP model	0.0	6854	0	0
proposed method	0.5	**128**	**6726**	**98.13**

of 15 consecutive frames, the RoI of the current frame will be classified as TP if it satisfies the two conditions: it overlaps with at least 7 RoIs of 7 neighboring frames, and their computed average confidence value is ≥0.5 (*avg*_*th*).

1) Evaluation of Positive Videos: Table II presents the results obtained on the 18 positive videos from CVC-ClinicVideoDB dataset. The maximum polyp detection capability of the two detector models including their FP models is obtained when the *score*_*th* = 0. However, when the *score*_*th* = 0, the num- ber of FPs is enormous i.e., low precision. In all cases, after applying the FP reduction method, we could significantly improve the precision and F1-score by keeping most of the TPs and eliminating most of the FPs. The reason that some TPs are classified as FPs is either that *avg*_*th* is less than 0.5 or the number of overlapping RoIs is less than 7. This TP degradation for the FP models is higher due to the fact that FP models produce softer predictions i.e., confidence of the detected polyps is smaller compared to the initial trained models. Compared to the initial Faster R-CNN and SSD models, the proposed method achieves the best overall performance by keeping a good balance between the sensitivity and precision (higher F1-score). This improvement is remarkably higher for FP models— ∼8% in the sensitivity and a little higher precision ∼(1%–3.5%).

2) Evaluation of Negative Videos: Table III presents the performance of the proposed method on 5 negative videos from ASU-Mayo Clinic. These 5 videos contain 6854 frames without polyps. When the confidence threshold of the RoI proposal network is 0.0, a RoI, which is obviously a FP, is provided for each frame. However, the proposed method can efficiently detect those FPs and outperform the counterpart models.

Based on the results of the initial Faster R-CNN and SSD, 68.02% and 84.01% of the proposed RoIs have a confidence value less 0.5, respectively. The proposed system is able to detect 16.24% and 9.64% (Faster R-CNN and SSD respectively) of those RoIs with confidence value more than 0.5. When the proposed method is applied to the FP models, the specificity can farther be improved and reaches close to 100%.

B. *Effect of Involving Previous or Future Frames Only*

To know how information from future and previous frames separately contribute to the performance increase, we conducted two extra experiments: 1) incorporating previous frames only, and 2) incorporating future frames only. Fig. 6 shows that incorporating previous frames enables the proposed method to remove FPs. More previous frames eliminate more FPs (i.e. better precision) whereas sensitivity decreases because some TPs will be removed in the final output detection. We obtained the same results when we incorporated future frames only (see Fig. 7). Again, the proposed method could not keep the sensitivity at the same level. Compared to Fig. 6, Fig. 7 makes sense because we are involving the same frames to make the final decision since the future frames become past frames dynamically. However, with the incorporation of both future and previous frames the method can detect less FPs and keep TPs, resulting in better F1_score (see Table II). We can conclude that involving

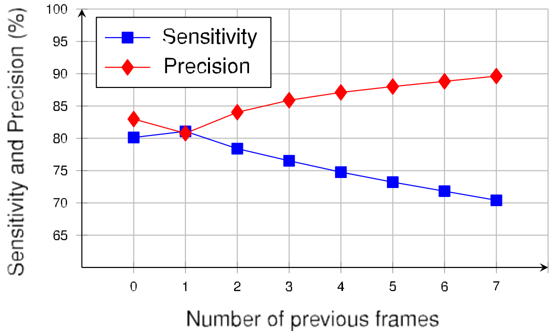


Fig. 6. Effect of involving only previous frames on the performance. Results were obtained on the 18 positive videos from CVC-ClinicVideoDB. With more previous frames, precision can be increased by removing FPs while sensitivity decreases because some TPs cannot be preserved.

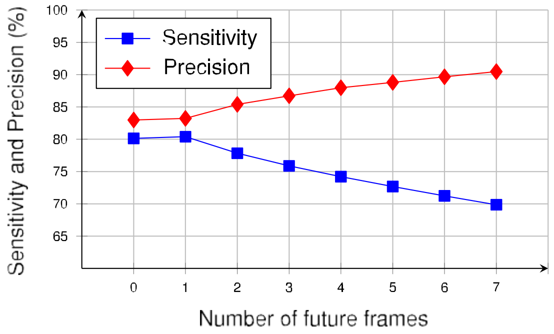


Fig. 7. Effect of involving future frames only on the performance. Results were obtained on the 18 positive videos from CVC-ClinicVideoDB. With more future frames, precision can be increased by removing FPs while sensitivity decreases because some TPs cannot be preserved.

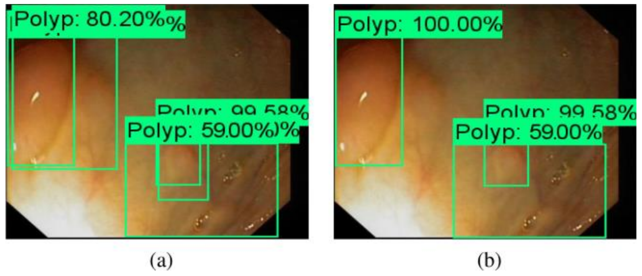


Fig. 8. An example where two detection outputs overlaid on the same regions. The redundant detection outputs with lower confidence values are eliminated by non-max suppression. (a) output detection before applying non-max suppression, (b) output detection after applying non-max suppression. Two RoIs eliminated by non-max suppression.

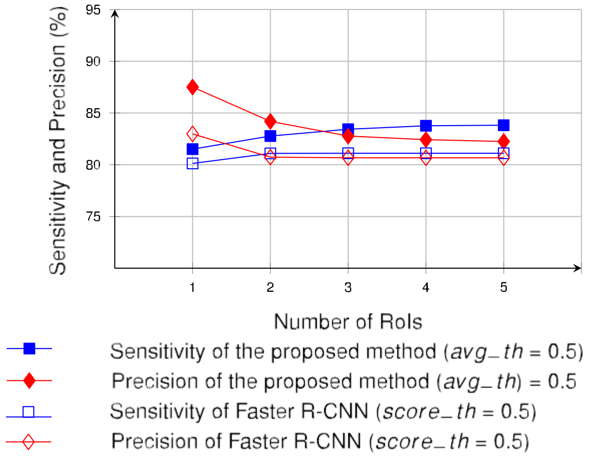


Fig. 9. Results obtained on the 18 positive videos from CVC-ClinicVideoDB dataset for multiple RoIs per frame scenarios using Faster R-CNN as the RoI proposal network in the first stage.

information from future and previous frames enables more reliable classification of FPs and TPs.

C. *Multiple RoIs per Frame*

Although in the positive test dataset there is no video that contains multiple polyps, multiple polyps on the colonoscopy frame can be possible. It is important for a CAD system to have the capability of detecting multiple polyps simultaneously. We conducted multiple RoIs per frame experiment for two purposes: 1) to confirm that the proposed method is robust to detect FPs even if several bounding boxes are provided, 2) to increase the detection capability in case the polyp is not bounded by the first box. That would confirm whether the model is suitable for multiple polyp detection task. If we set the detection output of the RoI proposal network to be n proposals, the top n RoIs will be returned. In this way, the model detection capability (sensitivity) increases whereas the precision decreases due to having a high number of FPs among these n proposals. It is necessary to run the optimization process again in order to obtain a new distance threshold value (dv). For example in case of 5 RoI proposals, we fixed $nf = 15$ *frames* and $dm = canberra$. The optimal dv changed from 0.65 to 0.55. We post-process the n proposed RoIs with non-max suppression to eliminate multiple redundant detections on top of the same polyp. In original Faster R-CNN and SSD [30], [31], Jaccard overlap thresholds of 0.7 and 0.45 were used, respectively. These thresholds might be optimal for object detection in natural images as there is possibility of having objects occluded by other objects. In colonoscopy, this possibility is rare, and we empirically noticed that the detectors would generate multiple redundant detections for the same polyp, and thus we fixed the Jaccard threshold at 0.25, see Fig. 8 as an example.

1) Evaluation of Positive Videos: We plotted the results of n RoIs proposal scenarios in Fig. 9. For sake of simplicity, we only show the results obtained when Faster R-CNN is used as the RoI proposal network. Similar results were obtained for FP models and SSD. Sensitivity slightly increases whereas precision degrades by involving more RoIs. However, both sensitivity and precision of the proposed method are improved compared to the counterpart models—initial models and FP models. This means our method can enhance the detection performance of both Faster R-CNN and SSD meta-architectures by integrating temporal information. Both sensitivity and precision tend to become constant after three RoIs. This is because the 100 RoIs generated by the first stage are sorted based on their confidence values. The deeper we go, the smaller the confidence value will be and the avg_th threshold condition eliminates those RoIs with low confidence values.

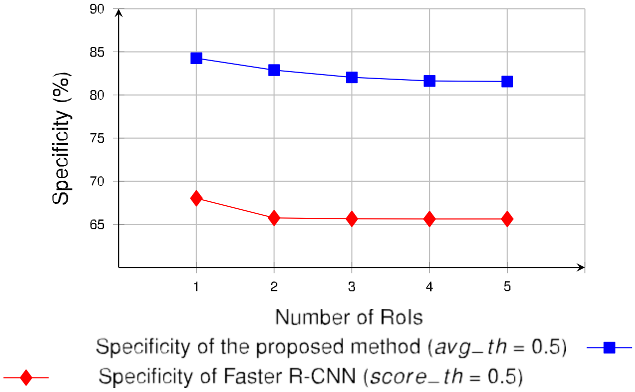


Fig. 10. Results obtained on the 5 negative videos from ASU-MAYO Clinic dataset for multiple RoIs per frame scenarios using Faster R-CNN as the RoI proposal network in the first stage.

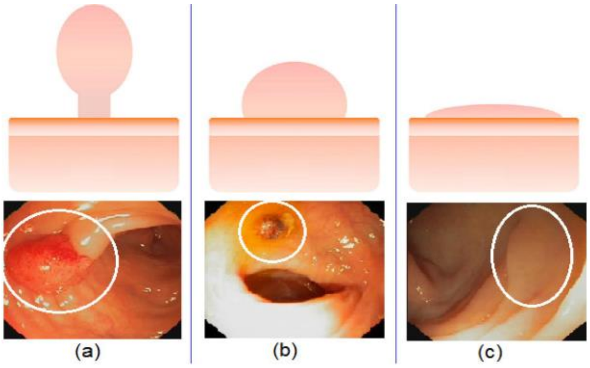
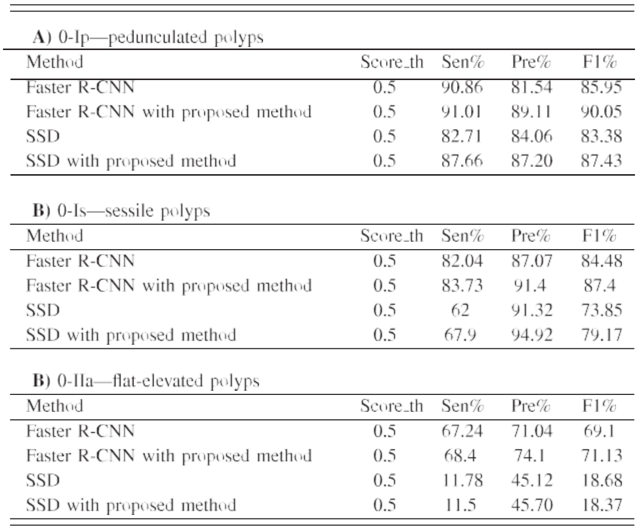


Fig. 11. Types of polyps in CVC-ClinicVideoDB. (a) 0-Ip— pedunculated polyp, (b) 0-Is—sessile polyp, (c) 0-IIa—flat-elevated polyp.

TABLE IV
PERFORMANCE EVALUATION OF FASTER R-CNN AND SSD IN DETECTING DIFFRENT TYPES OF POLYPS



A) 0-Ip—pedunculated polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	90.86	81.54	85.95
Faster R-CNN with proposed method	0.5	91.01	89.11	90.05
SSD	0.5	82.71	84.06	83.38
SSD with proposed method	0.5	87.66	87.20	87.43

B) 0-Is—sessile polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	82.04	87.07	84.48
Faster R-CNN with proposed method	0.5	83.73	91.4	87.4
SSD	0.5	62	91.32	73.85
SSD with proposed method	0.5	67.9	94.92	79.17

B) 0-IIa—flat-elevated polyps

Method	Score_th	Sen%	Pre%	F1%
Faster R-CNN	0.5	67.24	71.04	69.1
Faster R-CNN with proposed method	0.5	68.4	74.1	71.13
SSD	0.5	11.78	45.12	18.68
SSD with proposed method	0.5	11.5	45.70	18.37

2) Evaluation of Negative Videos: Fig. 10 shows that the proposed method is efficient to eliminate many of these FPs with confidence values ≥0.5 before displayed as the final detection. In Fig. 10, we again showed only the results obtained using Faster R-CNN as the RoI proposal network. We got similar results for the other models. For initial Faster R-CNN, the specificity is improved by 14.44% while for initial SSD this improvement was 8.77%. When applied on the FP models, the specificity of the proposed method was around 98% and still higher than the two FP models. When we take more RoIs into account we get slightly better sensitivity, and worse precision and specificity. These changes in the metrics will continue to repeat in the same manner if we take more than 5 RoIs. It will become unnecessary to conduct experiments for other scenarios.

D. *Performance Evaluation of Faster R-CNN and SSD*

It is important to evaluate the performance of Faster R-CNN and SSD in detecting different types of polyps. The polyps in the CVC-ClinicVideoDB dataset are categorized based on Paris classification by endoscopists. The statistics of this classification is given in [23]. Paris classification is based on morphology of polyps. This database contains only three types: 1) 0-Ip—pedunculated polyp in 1313 frames, 2) 0-Is—sessile polyp in 6633 frames, and 3) 0-IIa—flat-elevated polyp in 2079 frames. Fig. 11 illustrates the graphical representation of the three types of polyps with an example for each.

Table IV shows the detection capability of Faster R-CNN and SSD in detecting these three types of polyps. Both are able to detect all different types of polyps in at least a sequence of frames in all videos. Pedunculated polyps are the easiest type for both models. Faster R-CNN could detect 91.01% of pedunculated polyps whereas SSD could detect 87.66%. For sessile polyps, Faster R-CNN showed a better performance than SDD, with sensitivity of 83.73% and 67.9% respectively. For flat-elevated polyps SSD performed poor with sensitivity of 11.5% only while Faster R-CNN could detect 68.4% of them. These results show that Faster R-CNN is more powerful than SSD for flat polyps. In general, Faster R-CNN demonstrated better detection capability than SSD for all types of polyps. However, SSD is much faster than Faster R-CNN and meets real-time constraints. To evaluate the processing time, we use the Mean Processing Time (MPT)—the time needed for processing a frame and the time needed for displaying the results. On a standard PC with NVIDIA GeForce GTX1080i, MPT is 390 msec for Faster R-CNN while it is just 33 msec for SSD. The total MPT of the proposed method then becomes the MPT of the detectors (either 390 msec or 33 msec) plus the delay caused by the FP reduction unit (280 msec). The reason for these differences might be due to two factors: 1) the CNN feature extractor network of Faster R-CNN is much deeper, 2) there is an additional network (RPN) proposing RoIs in Faster R-CNN.

VI. DISCUSSION

Temporal information is essential to reduce the number of FPs in video sequences. Original Faster R-CNN and SSD meta-architectures are developed for object detection in still images

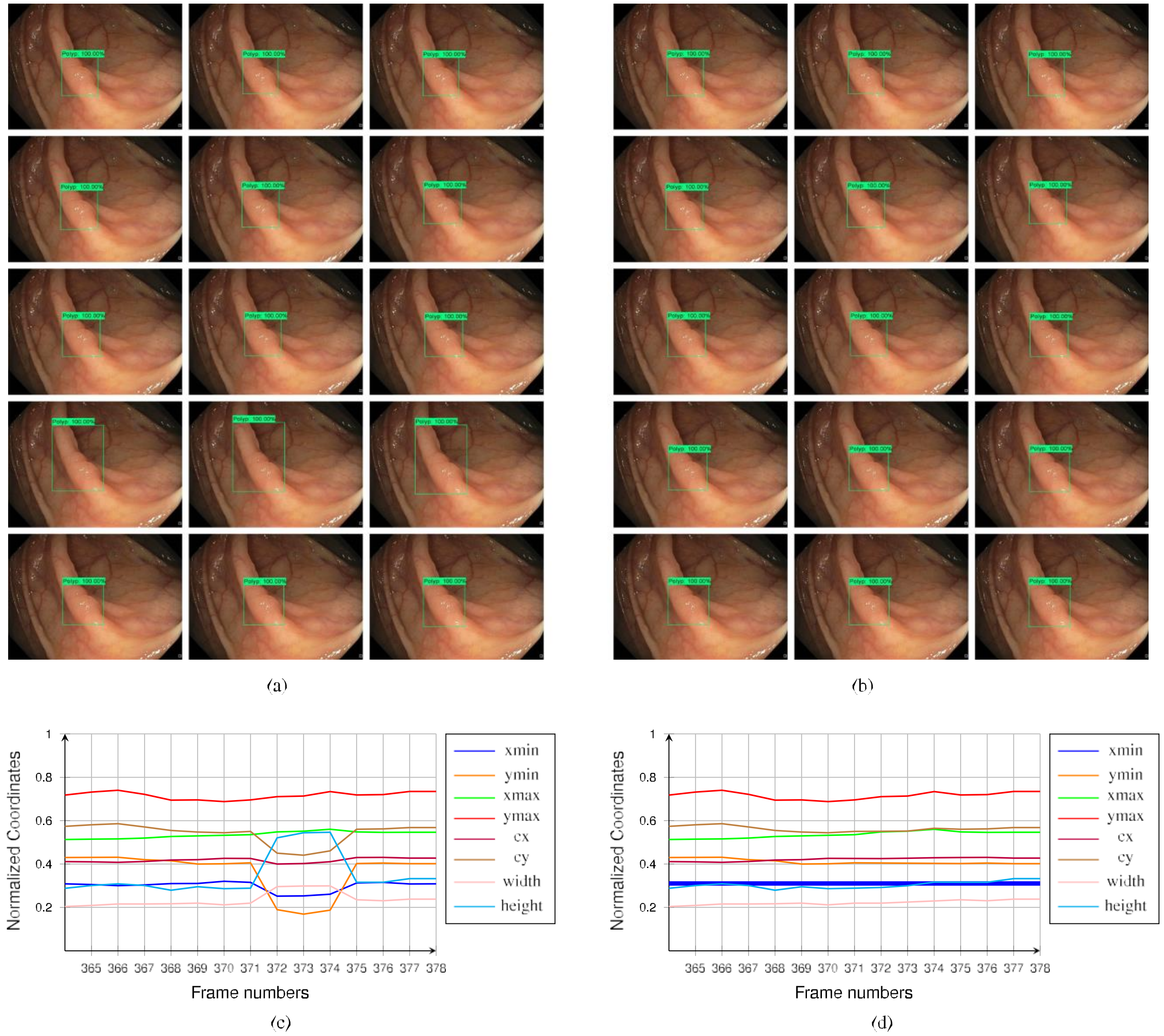


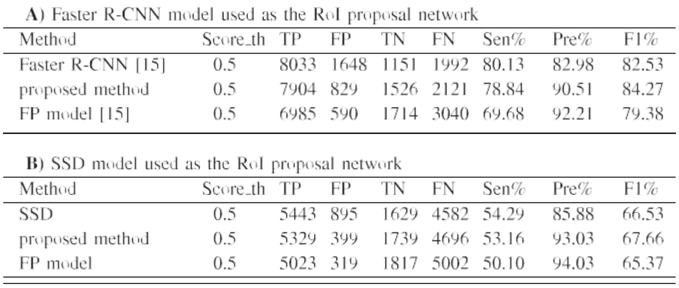
Fig. 12. Refining and smoothing the detection outputs in a sequence of frames starting from frame 364 (the top left frame in (a) and (b)) and ending at frame 378 (the bottom right frame in (a) and (b)). (a) Detection results before refining–see irregular detected bounding boxes in frames 372, 373, and 374, (b) Detection results after refining–see the corrected bounding boxes in frames 372, 373, and 374, (c) coordinates of the detected bounding boxes before refining, (d) coordinates of the detected bounding boxes after refining.

and do not have any mechanism to learn this important feature during training even if they are trained on video sequences. To improve their performance for polyp detection and make them more suitable for clinical usability, we integrated information from previous and future frames. The proposed scheme can be incorporated with any detector network for normal video detection applications. Usually, FPs are located in different positions in the neighboring frames, and their coordinates are irregular. The advantage of integratinginformation from future frames is to detect those irregularities with more robust and reliable decision-making and to estimate the changes in polyp position by a simple interpolation in order to detect missed polyps in inter-frames. The second advantage is to smoothen the detection output in the sequence by refining coordinates of those TP bounding boxes that are a little larger or smaller than those in the neighboring frames. In Fig. 12, even though the detections in frame 373, 374, and 375 are correct, the system recognizes them as abnormal relative to the detections in the consecutive frames and refines them using the same interpolation formula.

The main drawback of using future frames is that a small delay in displaying the detection outputs is introduced. The RoI proposal network generates RoIs for the last frame, but they will not be shown till the frame becomes the current frame—the frame in the middle of the sequence. In case of having 25 frames per second, this delay is just 280 msec. The main objective of the FP learning is to teach the detection models how FPs look like. Although this enhances both the precision and specificity, it degrades the sensitivity by a large ratio [15]. When we applied our FP reduction method over the results obtained by the initial Faster R-CNN and SSD (*score*–*th* = 0.5), we could improve

TABLE V

One RoI per Frame Scenario Results Obtained on 18 Positive Videos From CVC-ClinicVideoDB: in Each Sub-Table, the 1st Row Shows the Result of the Detectors With Score Threshold of 0.5, the 2nd Row Shows the Result of our Method Applied on the 1st Row Result, and the 3rd Row Shows the Results of FP Models for Comparison Purpose



A) Faster R-CNN model used as the RoI proposal network

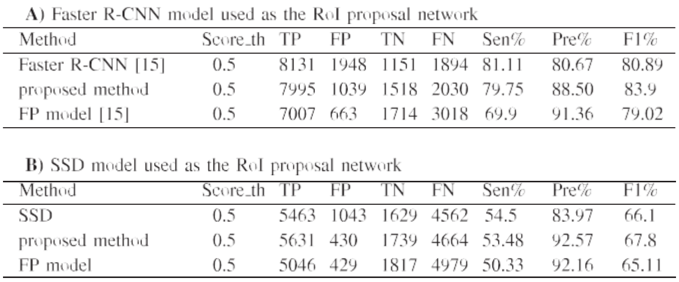
Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8033	1648	1151	1992	80.13	82.98	82.53
proposed method	0.5	7904	829	1526	2121	78.84	90.51	84.27
FP model [15]	0.5	6985	590	1714	3040	69.68	92.21	79.38

B) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5443	895	1629	4582	54.29	85.88	66.53
proposed method	0.5	5329	399	1739	4696	53.16	93.03	67.66
FP model	0.5	5023	319	1817	5002	50.10	94.03	65.37

TABLE VI

Five RoI per Frame Scenario Results Obtained on 18 Positive Videos From CVC-ClinicVideoDB Dataset: for More Details Please See the Caption of Table V



A) Faster R-CNN model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
Faster R-CNN [15]	0.5	8131	1948	1151	1894	81.11	80.67	80.89
proposed method	0.5	7995	1039	1518	2030	79.75	88.50	83.9
FP model [15]	0.5	7007	663	1714	3018	69.9	91.36	79.02

B) SSD model used as the RoI proposal network

Method	Score_th	TP	FP	TN	FN	Sen%	Pre%	F1%
SSD	0.5	5463	1043	1629	4562	54.5	83.97	66.1
proposed method	0.5	5631	430	1739	4664	53.48	92.57	67.8
FP model	0.5	5046	429	1817	4979	50.33	92.16	65.11

the precision by 7%–8% whereas the sensitivity got degraded by just 1%∼2%. From a clinical point of view, this balance is important and measured by the F1-score. As shown in Tables V and VI, the initial Faster R-CNN and SSD with the combination of our FP reduction unit have better sensitivity and thus better F1-score compared to their FP models.

Our method is similar to the methods proposed by Zhang *et al.* [17] and Yu *et al.* [18] in the way that all utilize temporal dependencies for better detection performance. However, Our method is developed to precisely eliminate FPs and keep/increase TPs. Unlike Zhang *et al.* [17], we used temporal information from future and previous frames. Future frames allowed us for better and more reliable decision making, and thus we were able to increase sensitivity, precision and specificity by keeping and increasing TPs and eliminating most of the FPs. Unlike Yu *et al.* [18], we used 2D-CNN for providing regions of polyp candidates and used 3D temporal information in a post processing unit to classify FPs from TPs. This makes our model less computationally and memory expensive compared to the 3D-CNN model in [18]. Unfortunately, due to licence problems we could not get our hands on the ground-truth of the ASU–Mayo Clinic test dataset to numerically compare all the three models in a table.

VII. Conclusions and Future Work

In this paper, we presented a novel polyp detection framework that can be used with any object detector method to integrate temporal information and increase the overall polyp detection performance in colonoscopy videos. The proposed scheme combines individual frame analysis and temporal video analysis to make the final decision in the current state. In particular, the proposed scheme benefits from the coordinates of the RoIs provided for a set of consecutive frames to measure the similarities and find detection irregularities and outliers. In addition, the proposed scheme is able to detect missed polyps and refine the detection output by incorporating some future frames. We validated our method on two state of the art convolutional neural network (CNN) based detectors, faster region based convolutional neural network (Faster R-CNN) and single shot multibox detector (SSD). Faster R-CNN is incorporated with the Inception-Resent for high detection performance, but low speed; SSD is incorporated with MobileNet for low detection performance, but real-time speed. Our experimental results showed that the two object detectors are missing the importance of Spatio-Temporal coherence feature for video sequence analysis and vulnerable to small changes, and thus they miss the same polyp within the inter-frames.

Only using the coordinates of the proposed RoIs to measure the similarities might not be sufficient to make the final detection decision. The possibility of incorporating additional features should be investigated to improve overall performance. It is important to find a mechanism in order to train the object detection models on video sequences to learn extra features such as motion estimation and variability of polyp appearance within a sequence of frames.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2018,” *Am. Cancer Soc.*, vol. 68, no. 1, pp. 7–30, 2018.

[2] M. Gschwantler *et al.*, “High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics,” *Eur. J. Gastroenterology Hepatology*, vol. 14, no. 2, pp. 183–188, 2002.

[3] M. Arnold, M. S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, “Global patterns and trends in colorectal cancer incidence and mortality,” *Gut*, vol. 66, pp. 683–691, 2016.

[4] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar, and P. D. Siersema, “Factors influencing the miss rate of polyps in a back-to-back colonoscopy study,” *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[5] L. Rabeneck, J. Souchek, and H. B. El-Serag, “Survival of colorectal cancer patients hospitalized in the veterans affairs health care system,” *Am. J. Gastroenterology*, vol. 98, no. 5, pp. 1186–1192, 2003.

[6] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, “Computer-aided tumor detection in endoscopic video using color wavelet features,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.

[7] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen, “Polyp detection in colonoscopy video using elliptical shape feature,” in *Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 2, pp. II-465–II-468.

[8] L. A. Alexandre, N. Nobre, and J. Casteleiro, “Color and position versus texture features for endoscopic polyp detection,” in *Proc. Int. Conf. BioMed. Eng. Informat.*, 2008, vol. 2, pp. 38–42.

[9] S. Ameling, S. Wirth, D. Paulus, G. Lacey, and F. Vilarino, “Texture-based polyp detection in colonoscopy,” in *Proc. Bildverarbeitung für die Medizin*, 2009, pp. 346–350.

[10] J. Bernal, J. Sánchez, and F. Vilarino, “Towards automatic polyp detection with a polyp appearance model,” *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[11] S. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim, “A colon video analysis framework for polyp detection,” *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1408–1418, May 2012.

[12] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians,” *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[13] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automated polyp detection in colonoscopy videos using shape and context information,” *IEEE Trans. Med. Imag.*, vol. 35, no. 2, pp. 630–644, Feb. 2016.

[14] J. Bernal *et al.*, “Comparative validation of polyp detection methods in video colonoscopy: Results from the MICCAI 2015 endoscopic vision challenge,” *IEEE Trans. Med. Imag.*, vol. 36, no. 6, pp. 1231–1249, Jun. 2017.

[15] Y. Shin, H. A. Qadir, L. Aabakken, J. Bergsland, and I. Balasingham, “Automatic colon polyp detection using region based deep CNN and post learning approaches,” *IEEE Access*, vol. 6, pp. 40950–40962, 2018.

[16] N. Tajbakhsh, S. R. Gurudu, and J. Liang, “Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks,” in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, 2015, pp. 79–83.

[17] R. Zhang, Y. Zheng, C. C. Y. Poon, D. Shen, and J. Y. W. Lau, “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker,” *Pattern Recognit.*, vol. 83, pp. 209–219, 2018.

[18] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos,” *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 65–75, Jan. 2017.

[19] R. Zhang *et al.*, “Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain,” *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 41–47, Jan. 2017.

[20] P. Brandao *et al.*, “Towards a computed-aided diagnosis system in colonoscopy: Automatic polyp segmentation using convolution neural networks,” *J. Med. Robot. Res.*, vol. 3, no. 2, 2018, Art. no. 1840002.

[21] N. Tajbakhsh, S. R Gurudu, and J. Liang, “System and methods for automatic polyp detection using convulutional neural networks,” US Patent App. 15/562 088, Mar. 15, 2018.

[22] Q. Angermann, A. Histace, and O. Romain, “Active learning for real time detection of polyps in videocolonoscopy,” *Procedia Comput. Sci.*, vol. 90, pp. 182–187, 2016.

[23] Q. Angermann *et al.*, “Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis,” in *Proc. Int. Workshop Comput. Assisted Robot. Endoscopy Clinical Image-Based Procedures*, 2017, pp. 29–41.

[24] S. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” *IEEE Conf. Comput. Vision Pattern Recog.*, Honolulu, HI, 2017, pp. 86–94.

[25] N. Narodytska and S. Kasiviswanathan, “Simple black-box adversarial attacks on deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, 2017, pp. 1310–1318.

[26] A. Nguyen, J. Yosinski, and J. Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 427–436.

[27] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in *Proc. ACM Asia Conf. Comput. Commun. Security*, 2017, pp. 506–519.

[28] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: A simple and accurate method to fool deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2574–2582.

[29] J. Su, D. V. Vargas, and S. Kouichi, “One pixel attack for fooling deep neural networks,” in *IEEE Trans. Evolutionary Comput.*, doi: 10.1109/TEVC.2019.2890858.

[30] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[31] W. Liu *et al.*, “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[32] Q. Dou *et al.*, “Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks,” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 779–788.

[34] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[35] R. Girshick, “Fast R-CNN,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, “Inception-v4, inception-resnet and the impact of residual connections on learning,” in *Proc. Nat. Conf. Artif. Intell.*, 2017, vol. 4, pp. 4278–4284.

[37] J. Huang *et al.*, “Speed/accuracy trade-offs for modern convolutional object detectors,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 4, 2017, pp. 3296–3297.

[38] A. G. Howard *et al.*, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv preprint arXiv:1704.04861.

[39] P. J. Davis, *Interpolation and Approximation*. Chelmsford, MA, USA: Courier Corporation, 1975.

[40] N. Tajbakhsh *et al.*, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[41] T. Lin *et al.*, “Microsoft coco: Common objects in context,” in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[42] T. Tieleman and G. Hinton, “Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude,” *COURSERA: Neural Networks Machine Learning*, vol. 4, 2012, pp. 26–30.

[43] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello, “Canberra distance on ranked lists,” in *Proc. Adv. Ranking NIPS Workshop*, 2009, pp. 22–27.