IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2019, accepted November 13, 2019, date of publication November 20, 2019,
date of current version December 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954675

A Framework With a Fully Convolutional Neural Network for Semi-Automatic Colon Polyp Annotation

HEMIN ALI QADIR[1,2,3], JOHANNES SOLHUSVIK[3], (Senior Member, IEEE), JACOB BERGSLAND[1], LARS AABAKKEN[4], AND ILANGKO BALASINGHAM[1,5], (Senior Member, IEEE)

[1]The Intervention Centre, Oslo University Hospital (OUS), 0372 Oslo, Norway
[2]OmniVision Technologies Norway AS, 0349 Oslo, Norway
[3]Department of Informatics, University of Oslo (UiO), 0373 Oslo, Norway
[4]Department of Transplantation, Faculty of Medicine, University of Oslo (UiO), 0372 Oslo, Norway
[5]Department of Electronic Systems at the Norwegian University of Science and Technology (NTNU), 7012 Trondheim, Norway

Corresponding author: Hemin Ali Qadir (hqadir2011@my.fit.edu)

This work was supported by the Research Council of Norway through the Industrial Ph.D. Project under Contract 271542/O30.

ABSTRACT Deep learning has delivered promising results for automatic polyp detection and segmentation. However, deep learning is known for being data-hungry, and its performance is correlated with the amount of available training data. The lack of large labeled polyp training images is one of the major obstacles in performance improvement of automatic polyp detection and segmentation. Labeling is typically performed by an endoscopist, who performs pixel-level annotation of polyps. Manual polyp labeling of a video sequence is difficult and time-consuming. We propose a semi-automatic annotation framework powered by a convolutional neural network (CNN) to speed up polyp annotation in video-based datasets. Our CNN network requires only ground-truth (manually annotated masks) of a few frames in a video for training and annotating the rest of the frames in a semi-supervised manner. To generate masks similar to the ground-truth masks, we use some pre and post-processing steps such as different data augmentation strategies, morphological operations, Fourier descriptors, and a second stage fine-tuning. We use Fourier coefficients of the ground-truth masks to select similar generated output masks. The results show that it is possible to 1) produce $\sim$ 96% of Dice similarity score between the polyp masks provided by clinicians and the masks generated by our framework, and 2) save clinicians time as they need to manually annotate only a few frames instead of annotating the entire video, frame-by-frame.

INDEX TERMS Colonoscopy, polyp segmentation, convolutional neural networks, semi-automatic, annotation, semi-supervised.

I. INTRODUCTION

Colorectal cancer (CRC) is the second and third most commonly diagnosed cancer in the world for females and males, respectively [1]. Most cases of CRC originate from small benign mucosal protrusions called adenomatous polyps. Over time, some of these polyps can turn into cancer if left untreated [2]. Colonoscopy is the preferred method for the detection and removal of such polyps, alternatively detecting early cancers when they can be successfully treated [3]. Colonoscopy is, however, operator dependent, and polyp miss-rate is reported around 22%-28% during colonoscopy [4].

Deep learning approaches, specifically convolutional neural networks (CNN), have demonstrated a strong performance for polyp detection and segmentation [5]–[12]. Not only do such deep models outperform traditional machine learning methods, but they also come with the benefit of not requiring difficult feature engineering. However, deep learning is a data-driven and data-hungry approach, i.e., its performance is highly correlated with the amount of available training data. The lack of large labeled polyp training images is one of the

The associate editor coordinating the review of this manuscript and approving it for publication was Zhao Zhang.

major obstacles in performance improvement of automatic polyp detection and segmentation [12]–[16]. Although there are some publicly available datasets (e.g. [17]–[21]), higher quality and a larger quantity of fully annotated datasets of polyp images and videos are highly desirable [14], [15]. Unlike a still frame dataset, a database of polyp videos can preserve temporal dependencies among frames. This temporal information is helpful to improve the performance of polyp detection [9]–[11]. Collecting and anonymizing polyp videos might not be as difficult as annotating them. Expert endoscopists are required to interpret colonoscopy videos and annotate them frame by frame. This process is time-consuming, and unnecessary work has to be repeated for the same polyp that appears in a sequence of neighboring frames. This might be one of the main obstacles of not realizing a large labeled database of polyp videos.

In this paper, we propose a framework powered by a new CNN based network to semi-supervisingly segment out polyp regions in video sequences and eliminate most of the unnecessary work needed for polyp annotation task. Our CNN has an encoder to extract hierarchical features from the input images, and multiple decoders (MDe) to restore the extracted features into a mask image. Hence, we name our network MDeNet. For each video, clinicians need to provide ground-truth of only a few numbers of frames. We use the manually annotated frames with their ground-truth to fine-tune a pre-train CNN, our proposed network. We also use the ground-truth masks as reference annotations to monitor outputs of the proposed framework. Based on these references, the proposed framework will generate masks for the rest of the remaining frames in the video.

II. RELATED WORK

There are many annotation tools [14] where an annotator has to draw polygons around objects by numerous clicks on the object boundary. Bernal *et al.* [14] used the datasets of polyps from the Gastrointestinal Image ANAlysis (GIANA) challenge[1] to qualitatively compare their labeling method with other similar and popular annotation tools. These tools are impractical for annotating video frames due to the massive manual workload in terms of the required number of clicks and time per frame.

Interactive segmentation methods for annotation aim at reducing human interactions to a few clicks, and thereby reducing the time costs required for each image. In a weakly supervised manner, annotators can select objects of interest by providing weak annotations such as strokes and bounding boxes [22]–[24]. The conventional interactive segmentation methods [25]–[27] typically look at low-level clues, such as colors, texture, etc. to segment the target object, leading to poor segmentation in cases of similar foreground and background appearances. Recently, deep learning has played an important role in the improvement of interactive segmentation techniques [22]–[24]. Although the output of deep learning-based interactive segmentation approaches looks much better than the conventional methods, they require substantial user interactions to produce satisfactory segmentation. This problem limits the use of those models for video annotation.

Semi-supervised video segmentation is another approach to annotate video frames in a more timely and efficient manner. In this approach, a segmentation model tries to provide annotations for the remaining frames of a video after it has been exposed to manual labels of a few frames of the same video. There are three trends to do this: propagation-based methods [28]–[34], appearance-based methods [35]–[37], and hybrid methods [38]. Propagation based methods leverage temporal coherence of object motion such as optical flow to propagate ground-truth labels from labeled to unlabeled frames. This approach seems to be vulnerable to temporal discontinuities like occlusions and rapid motion. It can also suffer from drifting once the propagation becomes unreliable [28]. To solve these problems, appearance-based methods have been proposed [35]–[37], in which a model learns the appearance of the target object from a set of given labeled frames, and then perform pixel-level detection of the target object at each frame. This approach seems to be vulnerable to appearance changes and object instances with similar appearances. Hybrid models aim to benefit from the advantages of both methods [38].

Our method falls in the line of hybrid research as we use temporal information among neighboring frames to strengthen an appearance-based model. Unlike other works [28]–[31], [35]–[38], which often train a model on manual labels of the first and/or the last frames, we recommend selecting k frames for manual labeling. That is because semi-automatic colonic polyp annotation in videos is challenging due to the complex environment of the inner lining of the colon (mucosa) and the existence of various polyp-like structures. In addition, when the endoscope moves in the colon, the appearance of the same polyp changes in neighboring frames. It will be difficult for a model to learn all the scene changes from the ground-truth of the frame where the targeted polyp first appears. We use the manually annotated ground-truth to fine-tune a pre-train CNN (our MDeNet) to learn the appearance changes of the target object from every interval period T. This is important for an annotation method to avoid generating unreliable masks and produce accurate segmentation so that they can be used as ground-truth images. Our novel algorithm provides an essential tool to reduce tedious manual labeling of video sequences. An annotator has to draw polygons around the target objects (polyps in our case) at the start, in some keyframes, and at the final frame.

III. METHODS

A. NETWORK ARCHITECTURE OF MDeNet

We would like our MDeNet to 1) accurately segment out the targeted polyps from the background with precise boundaries, 2) have a relatively small number of parameters so that it

[1] available at https://giana.grand-challenge.org/

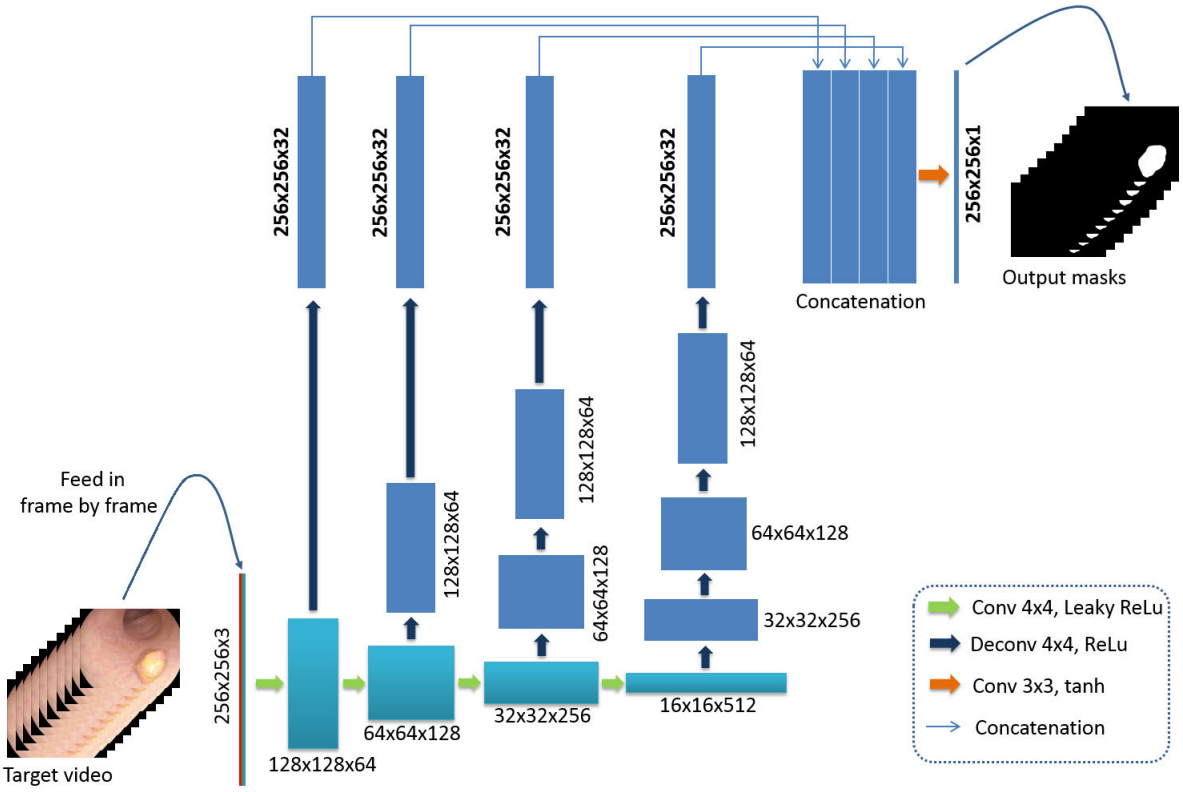


FIGURE 1. The network architecture of MDeNet. Every iteration, the network takes in a frame from the target video as the input RGB image of size 256 × 256 × 3, and generates a corresponding binary mask of size 256 × 256 × 1. The cyan boxes correspond to the encoder path, and the blue ones to the decoder paths. The resolution and the number of channels are denoted either at the bottom or next to the boxes such that the first two numbers are width and height, and the third is the number of channels.

can easily converge on a limited amount of manual annotation data, and have relatively fast inference times. Figure 1 illustrates the network architecture of MDeNet. It consists of an encoder and multiple paths of decoders. The encoder has four layers to extract different levels of features from the input image. At each layer of the encoder, there is a decoder to interpret the extracted features. In the encoder path, we lose some spatial information due to the contraction. We use multiple decoders to increase contextual and semantics information by utilizing the features from different scales. This step also increases the receptive field which helps to segment polyps of different sizes more precisely [39], [40]. We concatenate the outputs of the decoders by stacking them in a single layer. We apply a convolutional layer with *tanh* activation function on the concatenation layer to generate the output mask. This concatenation helps combine lower and higher levels of features in order to achieve accurate segmentation with satisfactory boundaries for the targeted objects.

A 4 × 4 unpadded convolution with stride 2 is applied for downsampling at each layer of the encoder path. Every convolutional layer is followed by a leaky rectified linear unit (Leaky ReLU) and batch normalization. We double the number of feature channels and halve the resolution at each down-sampling step. In each layer of the decoders, we upsample the feature maps by applying a 4 × 4 deconvolution with stride 2, each followed by a rectified linear unit (ReLU) and batch normalization. The decoder paths halve the number of feature channels and double the resolution. To generate binary polyp mask images, we concatenate the feature maps, which have the same dimensions of the input image, of the final layers of the decoder paths and apply a 3 × 3 padded convolution followed by *tanh* activation function.

The ground-truth of the training data is binary mask images, in which white pixels correspond to polyp pixels and black pixels correspond to the background. Xue *et al.* [41] showed that multi-scale L1 loss could force a CNN network to learn spatial relationships between pixels when features from multiple scales (i.e. multiple layers) are used to predict the output. Similarly, we predict the output binary masks from the concatenated feature maps decoded from multiple layers. Therefore, we choose the pixel-wise L1 loss as the objective function to update the network parameters in order to generate a precise boundary for the target polyps. Later, we evaluate other pixel-wise segmentation losses such as dice and cross-entropy losses. L1 loss computes the absolute error between the ground-truth mask X and generated output binary mask Y as follows:

$$\mathcal{L}^{\ell 1}(W) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{n}\sum_{i=1}^{n} \mid X_i - Y_i \mid,$$
$$Y = M(I; W), \tag{1}$$

where M is the CNN model, I is the input RGB image, W is the network parameters, m is the size of mini-batch, and n is the number of pixels.

B. PARENT MODEL

A large amount of training data is desired to train a CNN based network. If a limited number of training data is available, the network struggles to learn and find the global minima. It is our ambition to use as few labeled images as possible to reduce the amount of work required for manual polyp annotation. To learn the generic notion of polyp appearances, we use the binary masks of CVC-ColonDB dataset [18] (explained in Section IV-A) to pre-train the parameters of the CNN networks investigated in this study, including our MDeNet. We augment the images by rotating, zooming in & out, and shearing to increase the number of training images. For our MDeNet, we use Adam optimizer with a learning rate of 0.0002 and an exponential decay rate of 0.5 for 100 epochs. For the other networks, we use hyper-parameters recommended by the original papers for training. These pre-trained networks might fail to segment polyps from unseen images because they are unable to obtain generalization ability from this small training dataset. However, their parameters have some sort of knowledge of generic notion which helps the convergence of the networks when they are fine-tuned on the selected frames of the target videos.

C. FOURIER DESCRIPTORS

Polyp masks have a closed contour in the output binary image of the network. The closed contour can be approximated to an elliptical shape (see Figure 1). We use elliptic Fourier descriptors (FD) proposed by [42] for the characterization of closed contours. Even though the coefficients are invariant with the starting point, rotation, dilation, and translation, they contain precise information about the shape of the contour, and thus can be used for shape discrimination in binary images. Elliptic Fourier descriptors start from the chain code that approximates a continuous contour by numbering eight standardized line segments as follows

$$C = q_1 q_2 q_3 q_4 \ldots . q_K, \tag{2}$$

where each link q_i is an integer number between 0 and 7 oriented in the direction of $(\pi/4)q_i$. Fourier series expansion is appropriate for the x and y projections of the chain code because the code repeats on successive traversals of a closed contour. The truncated Fourier expansion for a closed counter can be written as

$$X_N = a_0 + \sum_{n=1}^{N} a_n \cos\frac{2n\pi t}{T} + b_n \sin\frac{2n\pi t}{T}, \tag{3}$$

$$Y_N = c_0 + \sum_{n=1}^{N} c_n \cos\frac{2n\pi t}{T} + c_n \sin\frac{2n\pi t}{T}. \tag{4}$$

N is the number of harmonics needed in the Fourier approximation. a_0 and c_0 are DC components and excluded from the features vector. a_n, b_n, c_n, and b_n are the coefficients which define the contour shape and can be calculated from the chain code as follows

$$a_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta x_p}{\Delta t_p}\left[\cos\frac{2n\pi t_p}{T} - \cos\frac{2n\pi t_{p-1}}{T}\right], \tag{5}$$

$$b_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta x_p}{\Delta t_p}\left[\sin\frac{2n\pi t_p}{T} - \sin\frac{2n\pi t_{p-1}}{T}\right], \tag{6}$$

$$c_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta y_p}{\Delta t_p}\left[\cos\frac{2n\pi t_p}{T} - \cos\frac{2n\pi t_{p-1}}{T}\right], \tag{7}$$

$$b_n = \frac{T}{2n^2\pi^2}\sum_{p=1}^{K}\frac{\Delta y_p}{\Delta t_p}\left[\sin\frac{2n\pi t_p}{T} - \sin\frac{2n\pi t_{p-1}}{T}\right]. \tag{8}$$

where t_p is the time required to traverse the first p links in the chain code, and x_p and y_p are, respectively, the projections on x and y of the first p links of the chain code.

D. PROCEDURE OF THE PROCESS

Figure 2 illustrates the entire procedure of the proposed framework, which consists of two trials. In the first trial, for each specific video, we initialize the network parameters from the parent model. We select a frame with a selection frequency of T in the target video V

$$V = \{f_1, f_2, f_3, f_4, \ldots\ldots\ldots, f_l\}. \tag{9}$$

We set the selection frequency to be $T = 50$, i.e. a frame is selected at every 50 consecutive frames. The selected frames which we call them reference frames F_r with their manual masks M_r, respectively, are

$$F_r = \{f_1, f_{50}, f_{100}, f_{150}, \ldots\ldots\ldots, f_l\}, \tag{10}$$

$$M_r = \{m_1, m_{50}, m_{100}, m_{150}, \ldots\ldots\ldots, m_l\}. \tag{11}$$

We always include the first and last frames in the set of the selected frames. We apply different augmentation techniques on the selected frames to improve the performance. We only apply those augmentation strategies that may simulate different scene variations in real colonoscopy videos. To remove imperfections at the inner and outer boundaries of the generated masks, we perform morphological closing followed by morphological opening using the same structuring element of size 5×5. The closing operation can fill some small holes that may appear inside the generated masks. We apply a morphological filling-hole operation to eliminate this artifact from the final output.

The results of the first trial may not be convenient and accepted as ground-truth images. The same polyp may be missed and producing irregular shapes is possible. We propose a second trial to enhance the results. We use shape information of the reference ground-truth masks M_r to collect more frames with their generated masks in the target video from the results of the first trial. We combine the reference frames and the collected frames to enlarge the training data for re-fine-tuning MDeNet. We perform a bidirectional scan

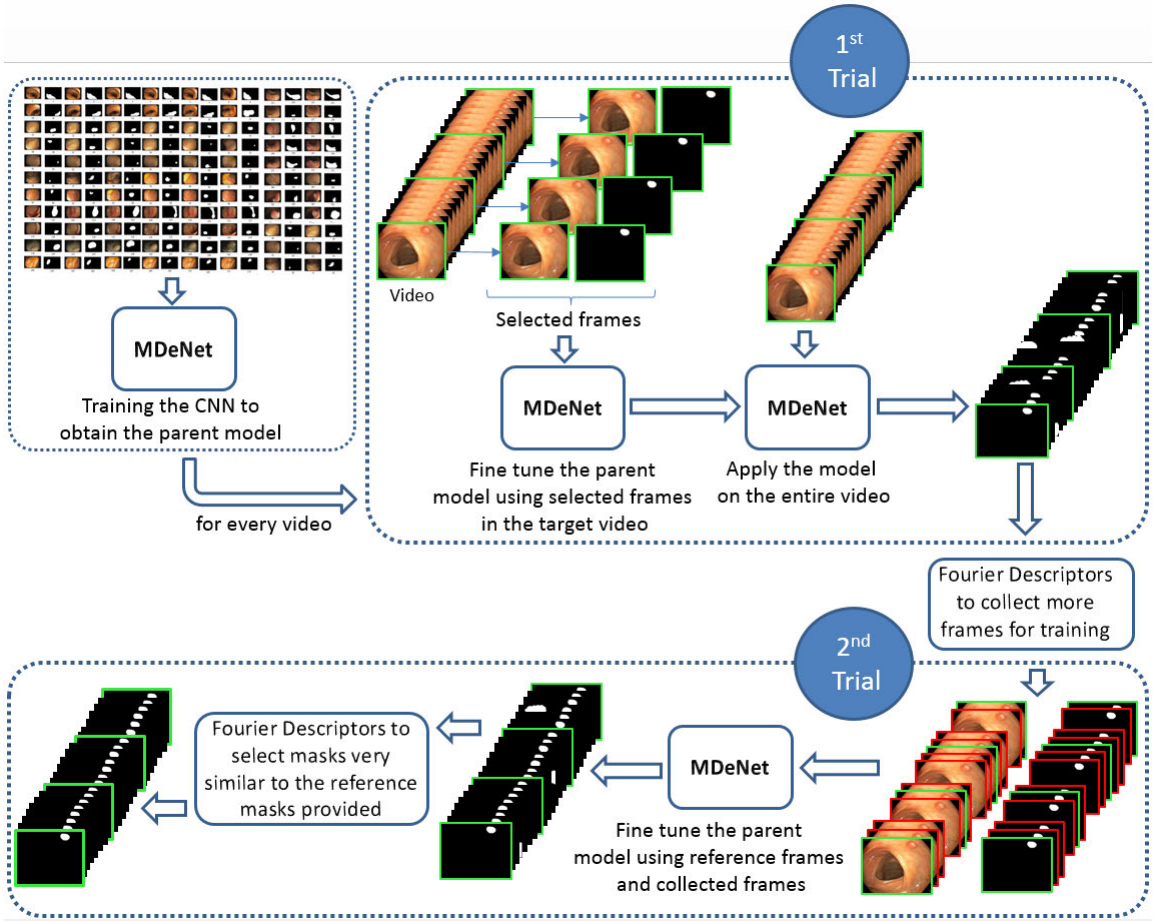


FIGURE 2. The entire procedure of the proposed method. MDeNet is pre-trained on a dataset of polyp images to obtain the parent model. The parent model is fine-tuned on a set of manually annotated reference frames (frames surrounded with green boxes) of the target video. The fine-tuned model is applied to the entire frames in the video. Fourier descriptor is used to eliminate irregular shapes generated by the model. More frames are collected (frames surrounded with red boxes) to further fine-tuning the parent model. The re-fine-tuned model is applied to all frames again. Fourier descriptor is applied to select only those generated masks similar to the reference masks.

on the generated masks from both sides of the reference images F_r to choose only those generated masks that are similar to the manual annotations M_r. We compute elliptic Fourier coefficients for every mask generated by the model and compare them with the coefficients of its corresponding reference mask using L_1-*norm*

$$\begin{aligned} L_1(m_i, m_g) &= \mid (FD(m_i) - FD(m_g) \mid, \\ m_i &\in M_r \\ i &= 1, 50, 100, ..., l, \\ for\ each\ i, \quad g &= i \pm 1, i \pm 2, i \pm 3, ..., i_{next/prev}. \end{aligned} \tag{12}$$

where m_i is the reference masks and m_g is the generated masks. In other words, we used Eq. 12 to take into account shape information and coherence information between the reference masks and the masks generated for the consecutive frames. Since Fourier descriptors are invariant to position, we robust the L_1-*norm* similarity measure by including the center of object mass. Again, we apply the same augmentations on the collected frames, fine-tune the model, and feed-in the entire target video to the retrained network. On the results of the second trial, we apply the same closing, opening, hole-filling, and bidirectional scan to eliminate irregular masks and imperfections.

IV. RESULTS AND DISCUSSION

A. DATASETS

We use two publicly available datasets: CVC-ColonDB dataset [18] which consists of 300 images of 15 unique polyps, and ASU-Mayo Clinic dataset [20] which consists of 38 fully annotated videos. We use CVC-ColonDB dataset to pre-train and initialize the parameters of the CNN networks in order to obtain their parent models as explained in Section III-B. Originally, the authors in [20] divided ASU-Mayo Clinic dataset into training and test subsets. They assigned 20 videos for the training phase and 18 videos for the test phase. We couldn't get access to the 18 videos assigned for the test phase due to licensing problems. Among the 20 videos assigned for the training phase, 10 are positive (with

TABLE 1. Performance improvement of the proposed framework in a step-wise manner.

Methods	MDeNet									
Original	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
+Rotation		✓	✓	✓	✓	✓	✓	✓	✓	✓
+Zoom-In			✓	✓	✓	✓	✓	✓	✓	✓
+Zoom-Out				✓	✓	✓	✓	✓	✓	✓
+Darkness					✓	✓	✓	✓	✓	✓
+Brightness						✓	✓	✓	✓	✓
+Closing & Opening							✓	✓	✓	✓
+Filling holes								✓	✓	✓
+Fourier Descriptor									✓	✓
+2nd trial										✓
Dice	0.649	0.745	0.765	0.778	0.783	0.793	0.820	0.822	0.854	0.946
improve by %	0	9.6	2	1.3	0.5	1	2.7	0.2	3.2	9.2
Jaccard	0.607	0.689	0.710	0.724	0.728	0.739	0.767	0.772	0.805	0.933
improve by %	0	8.2	2.1	1.4	0.4	1.1	2.8	0.5	3.3	12.8

polyps), and 10 negative (without polyps). In our test phase, we only need to use 10 positive training videos to evaluate the performance of the proposed framework. Although there exist some mis-labelings in the ground-truth images, this dataset is the only publicly available polyp dataset useful for quality assessment of the proposed annotation framework. This is because the polyp masks are polygon boundaries manually drawn by endoscopists. This enables us to compare the quality of annotations obtained by the proposed algorithm to the annotations provided by endoscopists.

B. EVALUATION METRICS

In order for any semi-automated annotation framework to be practically useful, it has to generate labels similar to the ground-truth provided by experts. In our case, we need to compute the overlap percentage between the polyp masks generated by the proposed method and manual reference masks drawn by endoscopists. We use two well-known overlap ratio measures: Jaccard index (also known as intersection over union, IoU), and Dice similarity score. Jaccard index computes the intersection of generated masks, A, and reference masks, B, divided by the size of their union as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \quad (13)$$

Similarly, Dice computes the intersection of generated masks, A, and reference masks, B, divided by the average size of A and B as follows:

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}. \quad (14)$$

The two metrics are sensitive to misplacement of the segmentation label, and that makes them very useful metrics for performance evaluation of the proposed method.

C. PERFORMANCE IMPROVEMENT

For each test video in the ASU-Mayo Clinic dataset, we noticed that 100 epochs for the first trial and 30 epochs for the second trial were enough to fine-tune the parent model. Table 1 shows the performance improvement of the proposed framework in a step-wise manner. With only the original reference frames as the training data, the proposed method could obtain 64.9% of Dice and 60.7% of Jaccard. When we increase the training data by applying different augmentation strategies, the performance increases gradually. We applied the following augmentations on the reference frames: 1) rotations by 90°, 180°, 270°, horizontal and vertical flips; 2) Zooming in and out by 5%, 10%, 15%, 20%, 25%, and 30%; 3) brightening and darkening by 25% and 50%.

With these augmentations, we could enhance 14.4% and 13.2% of Dice and Jaccard, respectively. Morphological closing and opening added 2.7% on Dice and 2.8% on Jaccard. The improvement by the filling-hole operation is small because MDeNet produced very small hole artifacts. Closing and opening operations cannot remove FP objects with irregular shapes which might be generated at random places. We applied Fourier descriptors to choose only those generated masks similar to the reference masks and remove irregular shapes in the output images. With this post-processing, we could improve Dice by 3.2% and Jaccard by 3.3%. Figure 3 illustrates a case where Fourier descriptors could successfully eliminate those irregular shapes generated by MDeNet. After the second trial was applied, Dice and Jaccard improved dramatically by %9.2 and %12.8, respectively. Figure 4 shows the final output results of three video sequences after applying the second trail and the post-processing techniques.

D. WHY THE PARENT MODEL?

As discussed in Section III-B, the parent model has some basic knowledge of the generic notion of polyp appearances, but it is unable to segment polyps from unseen video frames without fine-tuning it on several selected frames in the video (see Table 2). Figure 5 demonstrates that the parent model helps to speed up the fine-tuning progress. Without the parent model, the network needs more time to learn. The time needed for convergence differs for each video and depends on the number of available reference frames for training.

For some videos when selection frequency T was 50, the network without the pre-trained parameters could not even converge after training for more than ten thousand

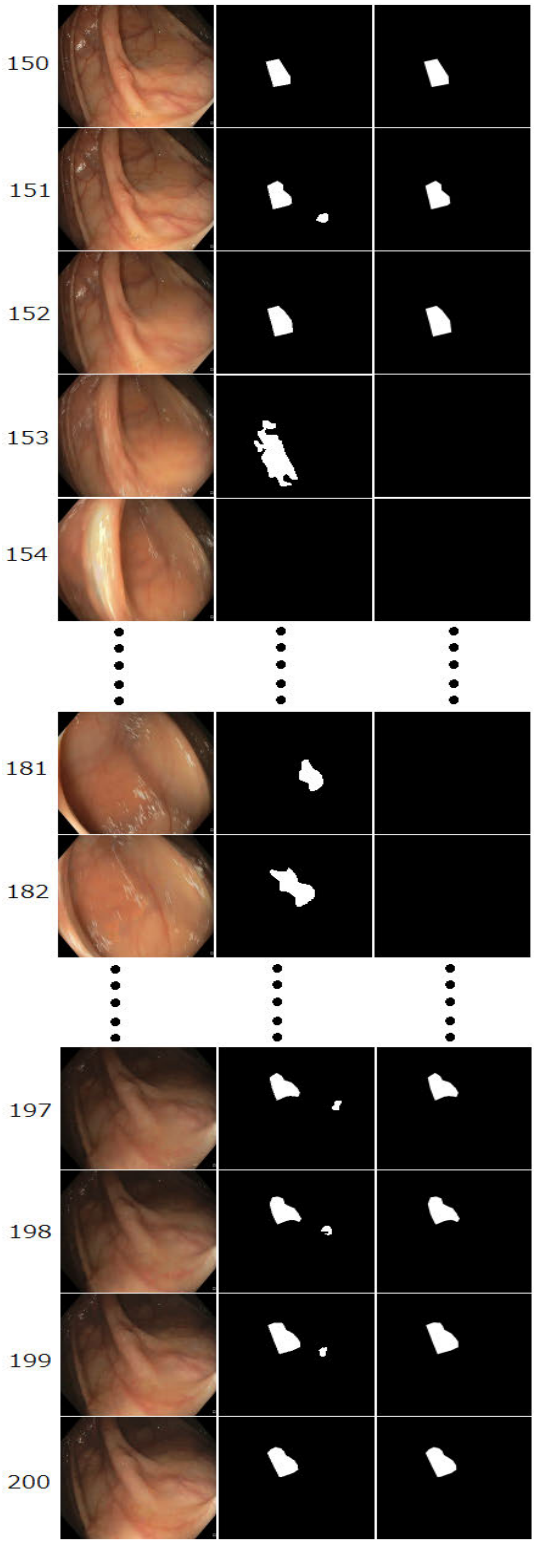


FIGURE 3. An example of using Fourier descriptors to remove irregular shapes. The numbers represent the frame sequence in the video, in which frames 150 and 200 are used as reference frames. Images in 1st column are the input RGB frames. Binary images in 2nd column are the output of the CNN network. Binary frames in 3rd column frames are the final output of the model after applying Fourier descriptors.

TABLE 2. Performance evaluation of MDeNet with and without parent model (pre-trained model).

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
Parent	0.381	0.351	-	-
Fine-tuning parent model	**0.854**	**0.805**	**0.946**	**0.933**
Training without parent model	0.727	0.693	0.804	0.792

TABLE 3. Effect of the number of reference frames on the performance of the framework.

selection frequency T	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
1	0.958	0.947	-	-
10	0.892	0.859	0.955	0.946
25	0.860	0.816	0.948	0.938
50	0.854	0.805	0.946	0.933
100	0.807	0.762	0.872	0.856

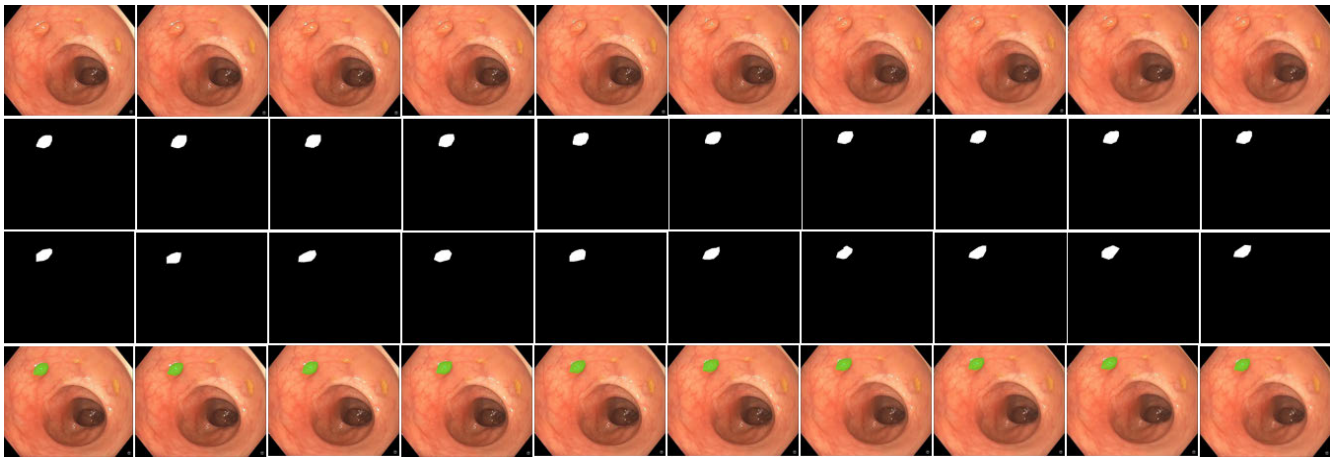
epochs. To guarantee network convergence the parent model becomes necessary. Table 2 shows that the results with the parent model are also better compared to the results without the parent model. That is because the model has never converged for two of the videos. In summary, the parent model helps the network converge in a very short time on a small selection frequency T, and improves the results for annotation.

E. IS IT OVER-FITTING?

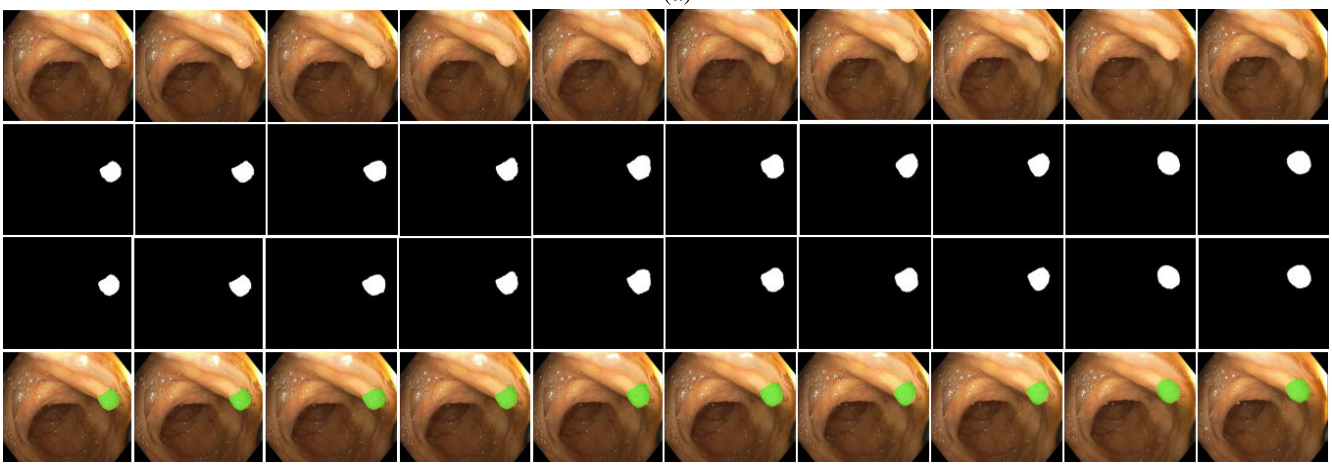
The way that we fine-tune the parent model to annotate the polyp in the target video may arise a question. One may ask "are we really trying to over-fit the network for the polyp in the target video?" To answer this question, we first fined-tuned the parent model for a polyp in one of the videos in ASU-Mayo Clinic dataset, and then applied it to annotate unseen polyps in other videos. Figure 6 shows that the fine-tuned model can only successfully annotate the polyp in the video used for fine-tuning, and fails to segment different polyps in other videos. Therefore, we can assume that the model gets over-fitted on the target video after the fine-tuning training.

F. EFFECT OF THE NUMBER OF REFERENCE FRAMES

In the previous experiments, we chose a frame at every 50 consecutive frames. Table 3 demonstrates how the performance improved when more frames were selected for the fine-tuning phase of the first trial. As shown in Table 3, selecting more frames for manual annotation could enhance the results of the first trial. However, we did not achieve a noticeable improvement in the performance of the second trial. This is due to the collection of extra training frames from the results of the first trial. When $T = 100$, the model was unable to obtain good results compared to the other cases. However, when $T = 50$, it seems to be enough for the framework to achieve results close to the results of $T = 10$.



(a)



(b)

FIGURE 4. The final output of the proposed framework for two target videos, each with a unique polyp. Each sub-figure (a and b) contains the following: the 1st row shows the input RGB frames, the 2nd row is the output binary masks generated by the model after applying all the post-processings, the 3rd row shows the ground-truth masks provided by clinicians, in the 4th row we overlay the output binary masks (2nd row) on top of the input RGB frames (1st row).

TABLE 4. Effect of using different loss functions for training MDeNet.

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
L1_Loss	**0.854**	**0.805**	**0.946**	**0.933**
Dice Loss	0.82	0.766	0.912	0.897
Entropy Loss	0.806	0.745	0.889	0.866

G. EFFECT OF USING DIFFERENT LOSS FUNCTIONS

In the previous experiments, we used L1 loss to train the models. In this experiment, we compare the performance of different pixel-wise loss functions, such as dice loss and binary cross-entropy loss, which are commonly used for image segmentation. Table 4 shows quantitative results of the three loss functions. The results confirm that L1 loss is able to generate better binary output masks from the concatenation layer decoded from multiple layers. We also surmise that this superior performance of L1 loss might be related to the reason that the model somehow tries to over-fit on the target polyps, and it seems that the L1 loss function is sufficient to help the model achieve this goal with better results.

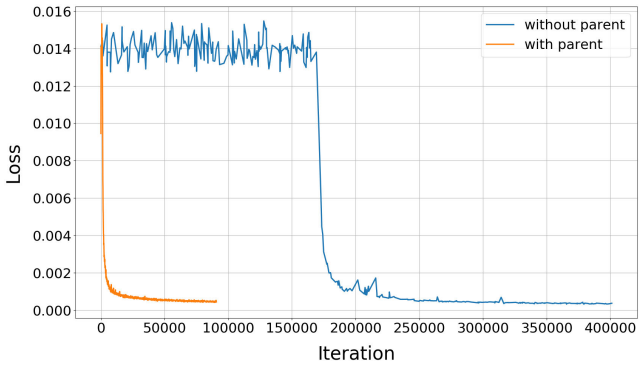


FIGURE 5. Fine-tuning progress for a video with and without the pre-trained parameters of the parent model.

H. PERFORMANCE COMPARISON OF MDeNet WITH OTHER CNN NETWORKS

In this experiment, we evaluate the performance of different well-known CNN architectures in our proposed framework shown in Figure 2. We replaced our CNN (MDeNet) with a

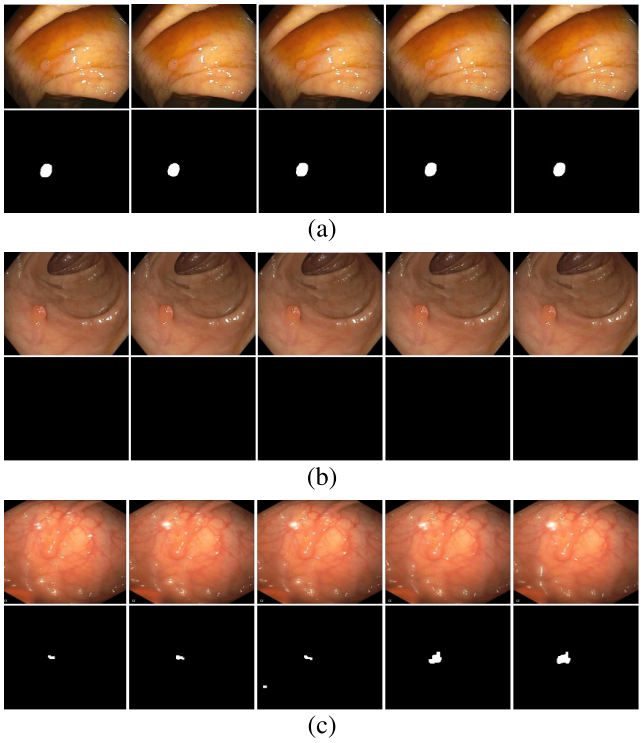


FIGURE 6. A case where the parent model was fine-tuned for the polyp appearing in video (a), and applied to annotate to unseen polyps in video (b) and (c). The fine-tuned models could successfully annotate the polyp in video (a) because it was already seen during fine-tuning. It failed to annotate the polyp in video (b). It could partly segment the polyp in video (c) because it seems to have some features of the polyp in video (a).

TABLE 5. Results of MDeNet compared with other CNN architectures used in the proposed framework.

Models	First Trial		Second Trial	
	Dice	Jaccard	Dice	Jaccard
MDeNet	**0.854**	**0.805**	**0.946**	**0.933**
U-Net	0.838	0.790	0.912	0.901
FCN	0.827	0.779	0.891	0.882
Mask R-CNN	0.812	0.761	0.876	0.818

fully convolutional neural network (FCN) [43], [44], a U-Net like network, and Mask R-CNN [45]. We used a U-Net architecture consisting of 8 layers in each its encoder and decoder paths. We used ResNet50 as the feature extractor network for Mask R-CNN. Compared to these CNNs, our MDeNet has less number of trainable parameters, meaning it has faster convergence and inference times. Table 5 shows that MDeNet has outperformed all the other three networks in both trials. This can be evidence for the ability of MDeNet to accurately segment out the target polyps from the background. Mask R-CNN is the state-of-the-art object segmentation method, however, it has performed poor for polyp annotation. There could be two reasons for this: 1) Mask R-CNN is developed for instance segmentation, not annotation, or 2) ResNet 50 is designed in such a way that much effort has been spent to prevent the model from over-fitting.

I. DISCUSSION

As noticed in the tables presented, in all cases the Dice similarity index is higher than the Jaccard index. Jaccard

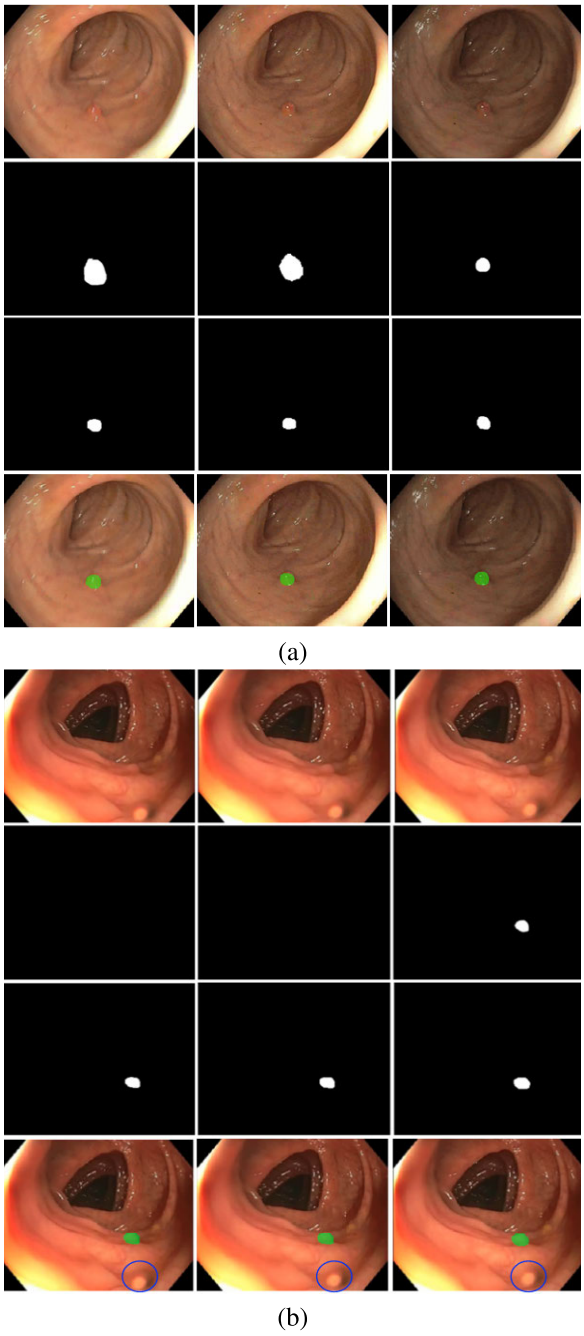


FIGURE 7. Two examples of manual annotation errors for the same polyps in three consecutive frames. Each sub-figure (a and b) contains the following: 1st row frames are the input RGB, binary images in the 2nd row are annotations provided by clinicians, and binary images in the 3rd row are the final output of the model, in the 4th row we overlay the output binary masks (3rd row) on top of the input RGB frames (1st row). *Note: The region bounded by the blue circle is an artifact from light reflection that looks like a polyp. This artifact can also be considered as an example of one of the challenges to differentiate between real and fake polyps when it comes to polyp detection and segmentation.*

is numerically more sensitive to mismatch when there is a reasonably strong overlap. Therefore, the Dice index is currently more popular than the Jaccard overlap ratio.

As shown in table 3, even when $T = 1$ we struggled to exceed 96% of Dice because the manual annotations by

clinicians in ASU-Mayo Clinic dataset are not free from human imperfections. Figure 7 illustrates two examples of manual errors in the test dataset. Figure 7.a shows that clinicians draw masks with different sizes for the same polyp in three consecutive frames whereas our model could give consistent annotations. Figure 7.b shows that clinicians missed the same polyp in two consecutive frames whereas the model was successful to nicely segment it from the background in all frames. This consistent segmentation is a clear advantage of using deep learning for qualitative annotation. Approximately 30 seconds to 1 minute is required to manually annotate a frame. With our framework and MDeNet, at least 2 hours can be saved for a video clip of 300 frames as we need clinicians to annotate only 6 frames to get satisfactory segmentation.

V. CONCLUSION AND FUTURE WORK

We proposed a semi-automatic framework for polyp annotations in video-based datasets. For this, we developed MDeNet, a convolutional neural network (CNN) based network, which can be trained on a few manually annotated frames and generate masks for the rest of the frames. The aim was to reduce the time spent on the unnecessary repeated work to annotate consecutive frames and thus speed up the annotation process. This framework has the potential for not only endoscopic image annotation but for other forms of medical image semi-automatic segmentation. The results showed that ground-truth images similar to the ones provided by clinicians can be achieved with only a limited number of manually annotated frames. For future work, we aim to develop an efficient key-frame selection algorithm to choose only those frames that identify abrupt changes in the target video. The goal will be to select a few frames as possible for manual annotation and still be able to achieve satisfactory results.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, Cancer J. Clin.*, vol. 68, no. 6, pp. 394–424, 2018.

[2] M. Gschwantler, S. Kriwanek, E. Langner, and B. Göritzer, C. Schrutka-Kölbl, E. Brownstone, H. Feichtinger, and W. Weiss, "High-grade dysplasia and invasive carcinoma in colorectal adenomas: A multivariate analysis of the impact of adenoma and patient characteristics," *Eur. J. Gastroenterol. Hepatol.*, vol. 14, no. 2, pp. 183–188, Feb. 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, no. 4, pp. 683–691, Apr. 2017.

[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] 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.

[6] P. Brandao, E. Mazomenos, G. Ciuti, R. Caliò, F. Bianchi, A. Menciassi, P. Dario, A. Koulaouzidis, A. Arezzo, and D. Stoyanov, "Fully convolutional neural networks for polyp segmentation in colonoscopy," *Proc. SPIE*, vol. 10134, Mar. 2017, Art. no. 101340F.

[7] L. Zhang, S. Dolwani, and X. Ye, "Automated polyp segmentation in colonoscopy frames using fully convolutional neural network and textons," in *Proc. Annu. Conf. Med. Image Understand. Anal.* Springer, 2017, pp. 707–717.

[8] 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.

[9] 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, Nov. 2018.

[10] 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.

[11] H. A. Qadir, I. Balasingham, J. Solhusvik, J. Bergsland, L. Aabakken, and Y. Shin, "Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video," *IEEE J. Biomed. Health Informat.*, to be published.

[12] H. A. Qadir, Y. Shin, J. Solhusvik, J. Bergsland, L. Aabakken, and I. Balasingham, "Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor cnn always perform better?" in *Proc. 13th Int. Symp. Med. Inf. Commun. Technol. (ISMICT)*, May 2019, pp. 1–6.

[13] A. Mohammed, S. Yildirim, I. Farup, M. Pedersen, and O. Hovde, "Y-Net: A deep convolutional neural network for polyp detection," Jun. 2018, *arXiv:1806.01907*. [Online]. Available: https://arxiv.org/abs/1806.01907

[14] J. Bernal, A. Histace, M. Masana, Q. Angermann, C. Sánchez-Montes, C. R. de Miguel, M. Hammami, A. García-Rodríguez, H. Córdova, O. Romain, G. Fernández-Esparrach, X. Dray, and F. J. Sánchez, "Gtcreator: A flexible annotation tool for image-based datasets," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 14, no. 2, pp. 191–201, Feb. 2019.

[15] V. de Almeida Thomaz, C. A. Sierra-Franco, and A. B. Raposo, "Training data enhancements for robust polyp segmentation in colonoscopy images," in *Proc. IEEE 32nd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jun. 2019, pp. 192–197.

[16] W. Chao, H. Manickavasagan, and S. G. Krishna, "Application of artificial intelligence in the detection and differentiation of colon polyps: A technical review for physicians," *Diagnostics*, vol. 9, no. 3, p. 99, Aug. 2019.

[17] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[18] J. Bernal and J. Sánchez, and F. Vilariño, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, Sep. 2012.

[19] J. Bernal and 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, Jul. 2015.

[20] 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.

[21] Q. Angermann, J. Bernal, C. Sánchez-Montes, M. Hammami, G. Fernández-Esparrach, X. Dray, O. Romain, F. J. Sánchez, and A. Histace, "Towards real-time polyp detection in colonoscopy videos: Adapting still frame-based methodologies for video sequences analysis," in *Computer Assisted and Robotic Endoscopy and Clinical Image-Based Procedures*. Cham, Switzerland: Springer, Sep. 2017, pp. 29–41.

[22] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 373–381.

[23] J. Liew, Y. Wei, W. Xiong, S.-H. Ong, and J. Feng, "Regional interactive image segmentation networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2746–2754.

[24] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5230–5238.

[25] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in nd images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Jul. 2001, pp. 105–112.

[26] B. L. Price, B. Morse, and S. Cohen, "Geodesic graph cut for interactive image segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3161–3168.

[27] V. Vezhnevets and V. Konouchine, "GrowCut: Interactive multi-label ND image segmentation by cellular automata," in *proc. Graphicon*, vol. 1, Jun. 2005, pp. 150–156.

[28] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," Dec. 2018, *arXiv:1812.01593*. [Online]. Available: https://arxiv.org/abs/1812.01593

[29] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.

[30] N. Märki, F. Perazzi, O. Wang, and A. Sorkine-Hornung, "Bilateral space video segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 743–751.

[31] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2663–2672.

[32] Z. Zhang, F. Li, L. Jie, J. Qin, L. Zhang, and S. Yan, "Robust adaptive embedded label propagation with weight learning for inductive classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3388–3403, Aug. 2018.

[33] Z. Zhang, M. Zhao, and T. W. S. Chow, "Graph based constrained semi-supervised learning framework via label propagation over adaptive neighborhood," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2362–2376, Sep. 2015.

[34] Z. Zhang, L. Jia, M. Zhao, G. Liu, M. Wang, and S. Yan, "Kernel-induced label propagation by mapping for semi-supervised classification," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 148–165, Jun. 2019.

[35] S. Caelles, K. Maninis, J. Pont-Tuset, and L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 221–230.

[36] K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, and L. Leal-Taixé, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," Sep. 2017, *arXiv:1709.06031*. [Online]. Available: https://arxiv.org/abs/1709.06031

[37] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2167–2176.

[38] S. Wug Oh, J.-Y. Lee, K. Sunkavalli, and S. Joo Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.

[39] P. O. Pinheiro, T. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 75–91.

[40] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[41] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang, "SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation," *Neuroinformatics*, vol. 16, nos. 3–4, pp. 383–392, 2018.

[42] F. P. Kuhl and C. R. Giardina, "Elliptic Fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, no. 3, pp. 236–258, 1982.

[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[44] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[45] K. He, G. Gkioxari, and P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

HEMIN ALI QADIR received the B.Sc. degree in electrical engineering from Salahaddin University-Erbil, Iraq, in 2009, and the M.Sc. degree in image processing from the Florida Institute of Technology, Melbourne, FL, USA, in 2013. He is currently pursuing the Industrial Ph.D. degree with OmniVision Technologies Norway AS, in collaboration with Oslo University Hospital (OUH) and the Department of Informatics, University of Oslo, Oslo, Norway. His research interests are image processing and computer vision, more specifically in medical and automotive applications. He is currently more engaged to apply deep learning techniques.

JOHANNES SOLHUSVIK received the Ph.D. degree in CCD and CMOS image sensor (CIS) design from ISAE, Toulouse, France. After which, he joined ABB Corporate Research, Norway. In 1999, he established Photobit (Norway) CIS Design Center, which was acquired by Micron Technologies Inc., in 2001. During this time, he also had a part-time position at NTNU, Trondheim, where he was teaching CIS design. From 2004 to 2006, he expatriated to Micron's CIS design HQ in the USA, where he managed design teams locally as well as remote teams in Japan, U.K., and Norway. He then repatriated to Norway to focus on CIS Research and Development and joined Aptina Norway, in 2009, where he served as a Fellow and the CTO of the Automotive BU. He joined OmniVision Norway, in 2012, as a General Manager and CIS Chip Architect. He currently holds a 10% position as an Associate Adjunct Professor at the University of Oslo, teaching CIS circuits and systems. He is a member of IISS' Board of Directors and has served multiple years as a TPC Member for IISW, ISSCC, and ESSCIRC.

JACOB BERGSLAND received the medical and Ph.D. degrees from Oslo University, in 1973 and 2011, respectively. After an internship in Norway, he moved to the USA for education in Surgery. He was a Specialist in general surgery, in 1981, and in cardiothoracic surgery, in 1983. He was the Director of the Cardiac Surgery, Buffalo VA Hospital, the Director of the Cardiac Transplantation Program, Buffalo General Hospital, the Director of the Center for Less Invasive Cardiac Surgery, a Clinical Associate Professor of Surgery, The State University of New York at Buffalo, an Initiator of the hospital partnership between Buffalo General Hospital and the Tuzla Medical Center, Bosnia, in 1995, and a Developer of the Cardiovascular Surgery and Cardiology in Bosnia and Herzegovina. He is currently a Researcher and a Co-Investigator with The Intervention Centre, Oslo University Hospital, the Medical Director of the BH Heart Centre, Tuzla BIH, and the Medical Director of Medical Device Company, Cardiomech AS.

LARS AABAKKEN received the medical degree from the Faculty of Medicine, Oslo, Norway, in 1986. His Ph.D. thesis was on the gastrointestinal side effects of non-steroidal, anti-in amatory drugs. He is currently an attending Gastroenterologist with the Oslo University Hospital, Oslo, involved in endoscopic procedures, EUS, and motility studies. He is also a Professor with the Department of Transplantation, Faculty of Medicine, University of Oslo.

ILANGKO BALASINGHAM received the M.Sc. and Ph.D. degrees in signal processing from the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, in 1993 and 1998, respectively. His master's thesis was with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, USA. From 1998 to 2002, he was a Research Engineer developing image and video streaming solutions for mobile handheld devices with Fast Search & Transfer ASA, Oslo, Norway, which is now a part of Microsoft Inc. He was appointed as a Professor of signal processing in medical applications with NTNU, in 2006. Since 2002, he has been with The Intervention Center, Oslo University Hospital, Oslo, as a Senior Research Scientist, where he is currently the Head of the Wireless Sensor Network Research Group. From 2016 to 2017, he was a Professor by courtesy with the Frontier Institute, Nagoya Institute of Technology, Japan. His research interests include super robust short-range communications for both the in-body and on-body sensors, body area sensor networks, microwave short-range sensing of vital signs, short-range localization and tracking mobile sensors, and nanoscale communication networks.