The following publication 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," in IEEE Journal of Biomedical and Health Informatics, vol. 21, no. 1, pp. 65-75, Jan. 2017 is available at https://doi.org/10.1109/JBHI.2016.2637004.

Integrating Online and Offline 3D Deep Learning for Automated Polyp Detection in Colonoscopy Videos

Lequan Yu*, *Student Member, IEEE,* Hao Chen*, *Student Member, IEEE,* Qi Dou, *Student Member, IEEE,* Jing Qin, *Member, IEEE,* and Pheng Ann Heng, *Senior Member, IEEE*

***Abstract*—Automated polyp detection in colonoscopy videos has been demonstrated to be a promising way for colorectal cancer (CRC) prevention and diagnosis. Traditional manual screening is time-consuming, operator-dependent and error-prone; hence, automated detection approach is highly demanded in clinical practice. However, automated polyp detection is very challenging due to high intra-class variations in polyp size, color, shape and texture and low inter-class variations between polyps and hard mimics. In this paper, we propose a novel offline and online 3D deep learning integration framework by leveraging the 3D fully convolutional network (3D-FCN) to tackle this challenging problem. Compared with previous methods employing hand-crafted features or 2D-CNNs, the 3D-FCN is capable of learning more representative spatio-temporal features from colonoscopy videos, and hence has more powerful discrimination capability. More importantly, we propose a novel online learning scheme to deal with the problem of limited training data by harnessing the specific information of an input video in the learning process. We integrate offline and online learning to effectively reduce the number of false positives generated by the offline network and further improve the detection performance. Extensive experiments on the dataset of *MICCAI 2015 Challenge on Polyp Detection* demonstrated the better performance of our method when compared with other competitors.**

***Index Terms*—Automated polyp detection, colonoscopy video, computer aided diagnosis, convolutional neural networks, deep learning**

I. Introduction

Colorectal cancer (CRC) is the second leading cause of cancer death in the United States and is estimated to have caused 49,190 deaths in 2016 according to American Cancer Society [1]. Since adenomatous polyps (adenocarcinomas) are most likely to develop into CRC, early and accurate detection of polyps from optical colonoscopy videos is of great significance for prevention and timely treatment of CRC. However, manual screening not only is laborious and time-consuming, but also heavily relies on clinical experience. It easily suffers from miss detection, which has been reported to be as high as 25% [2]. Missed polyps can lead to the late diagnosis of colon cancer with a low survival rate [3]. Hence, automated detection methods are highly desirable in clinical practice. However, automated detection of polyps from colonoscopy videos is very challenging due to high intra-class variations in polyp size, color, shape, texture and location as well as the low inter-class variations between polyps and hard mimics (e.g. colon walls, specular spots and air bubbles). Fig. 1 shows several examples of polyps and their mimics from colonoscopy videos.

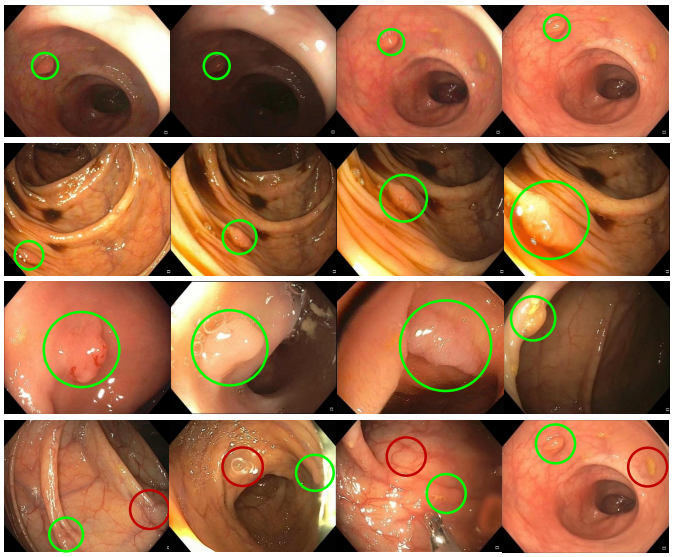


Fig. 1. The illustration of variations of polyps (green and red circles represent polyps and hard mimics). From the top to the bottom rows: the large color variation of the same polyp; the large size variation of the same polyp; the large shape variation among different polyps and low inter-class variation between polyps and hard mimics, respectively.

A. Related Work

Over the past few years, considerable efforts have been dedicated to developing efficient and robust approaches to automated polyp detection from colonoscopy videos. Most of these works detected polyps from general optical colonoscopy

Manuscript received October 8, 2016; revised November 25; accepted November 29. The work described in this paper was supported by the grants from the Research Grants Council of the Hong Kong Special Administrative Region (Project no. CUHK 14202514 and CUHK 14203115), the National Natural Science Foundation of China (Project No. 61233012) and Shenzhen Science and Technology Program (No. JCYJ20160429190300857).

* The first two authors contributed equally to this work.

L. Yu, H. Chen and Q. Dou are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China (e-mail: {lqyu, hchen}@cse.cuhk.edu.hk).

J. Qin is with the Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University, Hong Kong, China.

P. A. Heng is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China and also with Shenzhen Key Laboratory of Virtual Reality and Human Interaction Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China.

(OC) images while there were also some works detecting polyps from narrow-band imaging (NBI) colonoscopy data [4].

Some previous studies utilized polyps' color and texture information to design hand-crafted descriptors [5], [6], [7], [8]. For example, Karkanis *et al.* [5] employed color wavelet texture features as descriptors and combined sliding window strategy to detect polyps in colonoscopy images. Later, researchers proposed to utilize shape, intensity, edge and spatio-temporal features for automated polyp detection. For examples, Hwang *et al.* [9] adopted elliptical shape features to detect the presence of polyps; Bernal *et al.* [10], [11] presented a polyp region descriptor based on the depth of a valleys image and developed a region growing approach to locate polyps in colonoscopy images; Wang *et al.* [12], [13] utilized edge cross-section profiles for automated detection of protruding polyps; Ganz *et al.* [4] proposed an automated method to detect polyps in NBI colonoscopy data based on shape of polyps; Park *et al.* [14] employed the spatio-temporal features with the conditional random field model for automated polyp detection. Some methods combining two or more features have also been proposed to improve the detection performance [15], [16]. Tajbakhsh *et al.* [16] integrated the global geometric constraints and local intensity variation patterns to detect polyps. Although considerable advancements have been achieved, these methods still suffer from a low detection accuracy. The main reason is that the representation capability of hand-crafted features is quite limited to deal with the high intra-class variations of polyps and low inter-class variations between polyps and hard mimics.

Recently, deep convolutional neural networks (CNNs) with hierarchical feature learning capability trained on a large amount of training dataset have demonstrated state-of-the-art performance in many medical image analysis tasks, including classification [17], [18], [19], object detection [20], [21], [22], [23] and segmentation [24], [25], [26], [27], [28], [29]. As for automated polyp detection, some researchers also attempted to employ CNNs to handle this challenging task. For example, Tajbakhsh *et al.* [30] proposed a 2D-CNN method for polyp detection through taking the candidates selected by low-level hand-crafted features as input and utilizing an ensemble of 2D-CNNs to learn color, shape and temporal features of polyps. However, this method learned spatial and temporal features with different networks, which may somehow limit its discrimination capability. In this case, the rich spatio-temporal features of colonoscopy videos were not fully explored and harnessed.

While deep CNNs have achieved remarkable gains in medical image analysis tasks, most works focus on harnessing 2D-CNN to solve 2D image analysis problems. Recently, some researchers have proposed to employ 3D-CNN to deal with detection and segmentation tasks in volumetric medical data [31], [32], [33], [34]. These works demonstrated that 3D-CNN can achieve better performance than 2D-CNN and its variants when processing 3D medical data, as it can generate more discriminative features by taking full advantages of 3D spatial information. These works motivate us to explore the feasibility of 3D-CNN in endoscopic video processing, where we think it has great potential to generate representative spatio-temporal features for better outcomes. Actually, 3D-CNN has been proposed to recognize human actions from natural videos [35], [36], but we still face challenges to leverage it in medical video processing. One of the main concerns is that, compared to the large amount of training data for natural video processing tasks, the training data for medical applications are usually quite limited.

B. Our Contributions

In this paper, we propose an effective 3D fully convolutional network (3D-FCN) incorporated with a novel online and offline integration strategy for automated detection of polyps from colonoscopy videos. Different from the work reported in [30], our method learns spatio-temporal features simultaneously within a 3D-CNN framework to tackle the high intra-class and low inter-class variations of polyps. Besides, we further accelerate the detection progress by converting 3D-CNN into 3D-FCN without resorting to traditional time-consuming region proposal methods (e.g., sliding windows). More importantly, we propose a novel online learning scheme to deal with the problem of limited training data by integrating the specific information of an input video. By adaptively tuning the online network according to the specific testing video, this scheme can significantly reduce the number of polyp-like false positives. We evaluated our method on an open challenge dataset of *MICCAI 2015 Challenge on Polyp Detection*. Experimental results demonstrated that our method can achieve better performance than other competitors.

Our main contributions can be summarized as follows.

1) We propose an effective 3D-FCN to learn spatio-temporal feature representations for polyp detection from colonoscopy videos. Compared with previous methods based on hand-crafted features and 2D-CNNs, our method can more effectively tackle the large intra-class and low inter-class variations of polyps.
2) We propose a novel integrated framework with online and offline 3D representation learning to reduce the number of false positives and further improve the discrimination capability of our method for a specific video. This fusion learning strategy can remedy the deficiency of traditional CNNs in specificity caused by limited training data and improve their ability to handle cases with large variations.
3) Our method achieved the highest F1 and F2 score on the open challenge dataset of MICCAI 2015 Challenge on Polyp Detection.

The remainder of this paper is organized as follows. In Section II, we introduce the proposed 3D-FCN and the online learning scheme in detail. The experimental results are reported in Section III. We discussed some important issues relevant to this work in Section IV and conclusions are drawn in Section V.

II. METHOD

Fig. 2 shows the flowchart of the proposed framework, which integrates offline and online 3D representation learning by leveraging the 3D fully convolutional network (3D-FCN). An offline 3D-FCN (referred as *offline*-3D-Net) is first

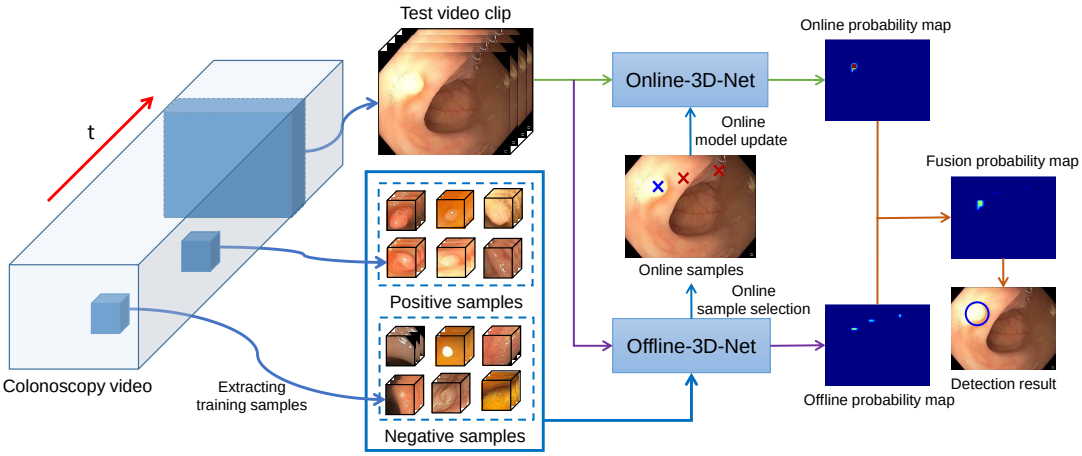


Fig. 2. The flowchart of the proposed online and offline 3D deep learning framework for automated polyp detection.

developed and exploited for learning spatio-temporal features from the training samples extracted from colonoscopy videos. Then, we incorporate an *online*-3D-Net, which is incrementally updated in the detection process for each input video, to effectively remove false positives generated by the *offline*-3D-Net. Finally, we fuse the outputs of these two networks to obtain the detection results.

A. 3D Fully Convolutional Networks

1) 3D Convolutional Neural Networks: While previous works on polyp detection pay more attention to the spatial features of polyps, we think the temporal information in colonoscopy videos also provides important clues for automated detection methods. Considering 3D-CNN can better encode spatio-temporal information in videos [36], we explore 3D-CNN to learn spatio-temporal features from colonoscopy videos for automated polyp detection. To the best of our knowledge, we are the first to employ 3D-CNN for endoscopic video analysis.

Typically, a 3D-CNN consists of 3D convolutional layers, 3D pooling layers, fully-connected layers and softmax layers. The 3D convolution and 3D pooling operations are performed in spatial and temporal dimensions. In addition, the outputs of 3D convolution and pooling are 3D feature volumes when the input is a video clip. In contrast, the outputs of 2D convolution and pooling are 2D feature maps even though the input is a video clip (taking multiple frames as multi-channels). To the end, 2D-CNN severely disregards the temporal information of colonoscopy videos through these 2D convolution and pooling operations. On the other hand, 3D-CNN can sufficiently preserve the temporal information of colonoscopy videos when extracting hierarchical features, hence it can effectively distinguish polyps from hard mimics such as specular spots and air bubbles [30] by taking full advantage of spatio-temporal information.

2) 3D Fully Convolutional Networks for Detection: By leveraging the representative spatio-temporal features learned from 3D-CNN, we can locate polyps from colonoscopy videos with sliding windows scheme through feeding cropped video sub-volumes into the 3D-CNN. However, this scheme is quite computationally expensive as thousands of candidate samples will be generated due to the high resolution and large frame number of colonoscopy videos. In addition, the polyps are sparsely distributed in the whole videos (most frames have only one or no polyps), making this traditional scheme inefficient and not applicable in clinical practice.

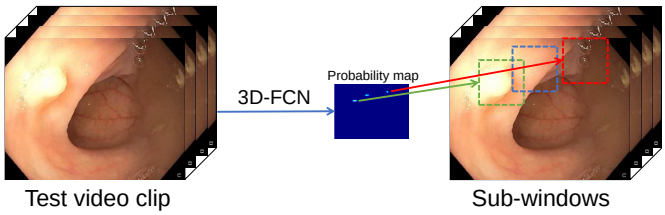


Fig. 3. The illustration of 3D-FCN which can generate the classification results of sub-windows within in one single forward propagation. Different boxes represent different sub-windows (overlapping cropped samples) and the size of the windows is the size of receptive field of 3D-FCN. Note that the size of probability map is smaller than that of the test video clip.

In this regard, we convert 3D-CNN to 3D fully convolutional network (3D-FCN) for fast detection by borrowing the fully convolutional concept in [37]. By converting the fully-connected layers in the 3D-CNN into convolutional layers, we obtain a 3D-FCN, which can take arbitrary-sized video clips as input and output corresponding probability maps . Compared with the sliding window scheme which repeatedly crops overlapping samples, our 3D-FCN can produce a probability map within one singe forward process. Each value in this probability map can be regarded as the network output of one sub-window (with same size of receptive field of 3D-FCN) in the original input video clips. Fig. 3 illustrates this process. Therefore, the 3D-FCN is inherently an accelerated variant of the traditional sliding window scheme. Note that due to the limitation of GPU memory, we input video clips with a specific length (16 frames in our experiments) instead of the whole video to the proposed 3D-FCN, and get one probability map within one single forward propagation. Our method is

quite different from the traditional sliding window method. In a sliding window approach, we need to repeatedly crop overlapping sub-volumes from one video clip and feed them to 3D-CNN to get the complete probability map of this video clip. While for our 3D-FCN, we only need to feed the whole video clip into the 3D-FCN and get the probability map of the whole video clip within a forward propagation directly. To the end, our method can reduce the redundant computations and accelerate detection compared to the traditional sliding window approach.

Due to the existence of down-sampling operations within the 3D-FCN, the dimensions of probability maps are reduced compared to the original input size and we need to determine the corresponding window location for each probability value. Supposing the spatial receptive field (the region in the original input video clip that influences the output probability value [38]) of the 3D-FCN is $r_w \times r_h$, the spatial down-sampling stride is $s_w \times s_h$ (the cumulative product of strides in convolutional and pooling layers; it is also the stride of sub-windows) and the spatial dimensions of input video clips are $w \times h$, the sizes of output probability maps ($w_p \times h_p$) can be calculated as:

$$w_p = \lceil \frac{w - r_w}{s_w} \rceil + 1,$$
$$h_p = \lceil \frac{h - r_h}{s_h} \rceil + 1. \quad (1)$$

We inverse the above equations to get the following index mapping equations:

$$x = \lceil \frac{r_w}{2} \rceil + s_w * (x_p - 1),$$
$$y = \lceil \frac{r_h}{2} \rceil + s_h * (y_p - 1), \quad (2)$$

where (x_p, y_p) and (x, y) represent the probability map index and the center location of the corresponding sub-window, respectively.

In order to detect polyps in the frame I_t (the t^{th} frame of a colonoscopy video), we first extract the neighboring frames centered at I_t to form a video clip with 16 frames and then feed the video clip into the proposed 3D-FCN to acquire a probability map. Finally we figure out the polyp locations by mapping the positions with probabilities above a threshold (0.8 in our experiments) in the probability map back to the input space according to Eq. (2).

B. Offline Representation Learning

1) Architecture of Offline 3D-FCN: The architecture of our proposed 3D-FCN used in offline representation learning is illustrated in Table I. Note that we here use a video clip with size of $102 \times 102 \times 16 \times 3$ (width×height×length×channel) as an illustration, but the 3D-FCN can take arbitrary-sized video clips as input. Previous studies [39], [40] have shown that small convolution kernels are more effective compared to the counterpart of large kernels with more discrimination capability while less computation parameters. For example, a stack of three $3 \times 3 \times 3$ convolutional kernels has an effective receptive field of $7 \times 7 \times 7$ but the stacked layers incorporate three non-linear rectification layers instead of a single one, which makes the decision function more discriminative [39]; assuming that both the input and output of the three-layer $3 \times 3 \times 3$ convolution stack have C channels, the three-layer stack has $3\times (3^3C^2) = 81C^2$ weights while the single 7×7×7 convolution kernel requires $7^3C^2 = 343C^2$ parameters. Hence stacked small kernels have less parameters and are more computationally efficient. We introduce this finding in our implementation of 3D convolutional networks by using small convolution kernels with size of 3×3×3 (spatial width×spatial height×temporal depth) in convolutional layers.

Overall, our network consists of 6 conventional convolutional layers (Conv) with size of 3×3×3 and each of them is followed by a rectified linear unit (ReLU) [41] as an activation function. We also add 4 max-pooling layers (Pool) between these convolutional layers to increase the receptive field and reduce the feature volume size. After each pooling layer, we double the number of feature volumes to preserve the necessary information. There are 2 converted convolutional layers (Conv5 and Conv6) followed by layer Pool4. These two convolutional layers are converted from fully-connected layers and can allow our network to take arbitrary-sized input.

TABLE I
THE ARCHITECTURE OF THE PROPOSED OFFLINE 3D-FCN (ARCH I).

Layer	Feature maps	Kernel size	Stride
Input	$102\times102\times16\times3$	-	-
Conv1a	$100\times100\times14\times64$	$3\times3\times3$	$1\times1\times1$
Pool1	$50\times50\times14\times64$	$2\times2\times1$	$2\times2\times1$
Conv2a	$48\times48\times12\times128$	$3\times3\times3$	$1\times1\times1$
Pool2	$24\times24\times12\times128$	$2\times2\times1$	$2\times2\times1$
Conv3a	$22\times22\times10\times256$	$3\times3\times3$	$1\times1\times1$
Conv3b	$20\times20\times8\times256$	$3\times3\times3$	$1\times1\times1$
Pool3	$10\times10\times8\times256$	$2\times2\times1$	$2\times2\times1$
Conv4a	$8\times8\times6\times512$	$3\times3\times3$	$1\times1\times1$
Conv4b	$6\times6\times4\times512$	$3\times3\times3$	$1\times1\times1$
Pool4	$3\times3\times2\times512$	$2\times2\times2$	$2\times2\times2$
Conv5	$1\times1\times1\times1024$	$3\times3\times2$	$1\times1\times1$
Conv6	$1\times1\times1\times2$	$1\times1\times1$	$1\times1\times1$

Note: Conv5 and Conv6 are converted from fully-connected layers.

2) Offline Model Training: We train the offline 3D-FCN (*offline*-3D-Net) using the cropped sub-volumes. As the ground truth of training colonoscopy videos are pixel-level annotated polyp masks, we use the following strategy to construct offline training samples. Given a polyp mask, we first calculate the centroid of this mask as the polyp location. Then a positive training sub-volume with size of 102×102×16×3 (the 3D-FCN will output one probability value) is cropped centered at the calculated polyp location. And the negative training sub-volumes are randomly cropped in the colonoscopy videos with no overlap with the positive training sub-volumes.

Training a deep CNN from scratch, i.e., the weights of networks are randomly initialized, is difficult because this manner requires a large amount of training samples. In addition, we need more data when training 3D networks because 3D networks have more parameters than 2D networks. However, the insufficiency of training data is a well-known problem of harnessing deep learning techniques in medical image computing. For example, there are only 10 videos containing polyps in our training data (see Section III-A for more detail). The limited training dataset would easily lead to overfitting problem when training deep networks. In order to partly tackle

the insufficiency of training data, we use the transfer learning technique [20] following previous works [18], [20], [21]. Specifically, we employ a pre-trained network [36] (trained on a large-scale video data Sport-1M [42], which contains 1.1 million sports videos) to initialize the weights of our 3D-FCN. Next we fine-tune our network with backpropagation method using the training sub-volumes.

C. Online Representation Learning and Model Fusion

By harnessing the spatio-temporal feature representations from colonoscopy videos, the *offline*-3D-Net can achieve good performance on polyp detection. However, due to the large variations across different videos, the *offline*-3D-Net trained in limited video clips may still output some polyp-like false positives. In our experiments, we observe that these polyp-like false positives are video-specific; the false positives in the same video are similar but the false positives in different videos are different. In this case, if the network can learn to discriminate specific false positives from each video, it can efficiently improve the precision performance. Based on above observation and consideration, we propose an online representation learning scheme to further improve the detection performance. More specifically, we train a specific online network (referred as *online*-3D-Net) for each testing video with online extracted samples from this video. This scheme can compensate the *offline*-3D-Net's inadequacy in discrimination capability caused by the gap between the large variations of polyps across different videos and the limited training dataset. Through online representation learning regarding a specific video, the *online*-3D-Net can leverage the specific information derived from this video and thus reduce the number of false positives.

1) Online Sample Selection: The key step of online representation learning is the selection of training samples, which should be representative for training *online*-3D-Net to enhance its capability of distinguishing polyps from hard mimics. We extract the online samples according to the results obtained from the offline network. When extracting the online training samples centered at I_t, we first generate three probability maps P_{ij}^{t-1}, P_{ij}^{t} and P_{ij}^{t+1} using *offline*-3D-Net and the video clips centered at frame I_{t-1}, I_t and I_{t+1}. Then we compare the three probability maps with a probability threshold P_o and obtain the positive ($\forall\tau \in \{t-1, t, t+1\}, P_{ij}^{\tau} > P_o$) and negative ($\exists\tau \in \{t-1, t, t+1\}, P_o - 0.2 < P_{ij}^{\tau} \leq P_o$) probability indexes (i, j). Next, the positive and negative positions corresponding to positive and negative indexes are localized based on Eq. (2). Finally, we extract the positive training samples from the localized positive positions while the negative samples consist of two parts: the samples selected from the localized negative positions and the samples drawn randomly without overlapping with the extracted positive samples. As the first part of negative samples have relative high probability values, adding these hard negative samples can enhance online model's capability of distinguishing polyps from polyp-like false positives. We employ the above strategy to extract online training samples from each frame of this video. The parameter P_o can be used to adjust the number of positive and negative training samples and we set it through cross-validation using the 10 colonoscopy videos containing polyps in our experiments. We show some extracted training samples for training the *online*-3D-Net in Fig. 4. It is observed that our online selection strategy can effectively extract polyp-like false positives as negative samples and hence can improve the capability of the online model to combat the hard mimics of polyps.

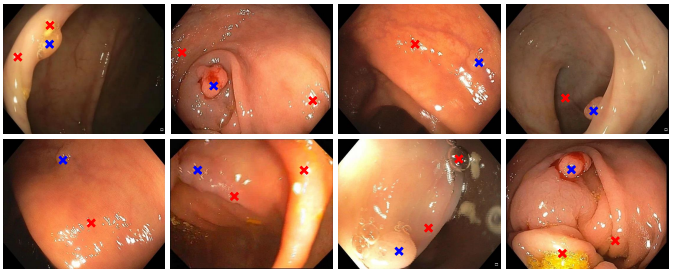


Fig. 4. Examples of extracted online training samples. The blue and red crosses represent positive and negative samples, respectively.

2) Online Model Learning: The online model is also implemented based on 3D-FCN and adopts the same architecture as the offline network. Considering the limited online training samples, we train the online network based on the offline network instead of training it from scratch. We use the weights of *offline*-3D-Net to initialize each *online*-3D-Net's weights and update its weights with backpropagation using online training samples extracted from this video. The online update is performed incrementally and in 60 frames interval using training samples extracted from previous frames. Note that only the last three convolutional layers of the online network are updated while the weights of previous convolutional layers are fixed throughout online model updating. This scheme is not only computationally efficient, but can avoid overfitting by fixing the video-independent spatio-temporal features extracted from previous convolutional layers.

3) Model Fusion: As we mentioned above, the main purpose of online network is to remove the specific polyp-like false positives detected by offline network and further improve the detection performance. To do this, we combine the outputs of offline network and online network to get final polyp detection results. Note that in online sample selection process, the probability maps generated by *offline*-3D-Net may not be correct if the training samples are contaminated by noise examples (though not common as observed from our experiments). In order to make our model more robust, we set a threshold T_s to bound the influence from *online*-3D-Net and calculate the final probability P_{ij} as:

$$P_{ij} = \begin{cases} \frac{P_{ij}^{\text{off}} + P_{ij}^{\text{on}}}{2} & |P_{ij}^{\text{on}} - P_{ij}^{\text{off}}| \leq T_s \\ P_{ij}^{\text{off}} & |P_{ij}^{\text{on}} - P_{ij}^{\text{off}}| > T_s \end{cases} \tag{3}$$

where P_{ij}^{off} and P_{ij}^{on} (i and j are the indexes of the probability maps) are the predicted probabilities of *offline*-3D-Net and *online*-3D-Net, respectively. If the absolute difference between these two outputs is greater than T_s, the online output is discarded and we only use the offline network result; otherwise, we use the average result as the final prediction. The T_s is set

as 0.3 in our experiments through cross-validation using the 10 colonoscopy videos containing polyps.

4) Complete Detection Flow of Our Method: We first train an *offline*-3D-Net using all the training sub-volumes and initialize an *online*-3D-FCN with the same weights of *offline*-3D-FCN for each testing video. Next we process the testing videos frame-by-frame by generating probability maps for each frame. When processing frame I_t, we extract a video clip with 16 frames (from I_{t-7} to I_{t+8}) and feed this video clips to *offline*-3D-Net and *online*-3D-Net to generate the offline and online probability maps. We fuse these two probability maps using Eq. 3 and generate the polyp locations using the steps described in Section II-A2. At the same time, we extract the online training samples and update the *online*-3D-FCN using the strategies in Section II-C1 and II-C2.

D. System Implementation

The proposed framework was implemented with C++ and Matlab under the open source deep learning library of Caffe [43] using a standard PC with a 2.60GHz Intel(R) Xeon(R) E5-2650 CPU and a NVIDIA GeForce GTX TITAN X GPU. The offline network was trained with standard backpropagation using stochastic gradient descend method (batch size=16, momentum=0.9, weight decay=0.005, the learning rate was set as 0.0005 initially and decreased by a factor of 10 every 4000 iterations). We updated the online network for 50 iterations with the same batch size, momentum and weight decay with the offline network, but set the learning rate for the last three convolutional layers as 0.001 for fast learning the specific information from testing videos. The parameter P_o in online sample selection was set as 0.8. Generally, it took 0.25 seconds to process one frame only using offline model; it took 1.23 seconds to process one frame using fusion model and about half of the time was spent in online network updating procedure.

III. Experiments and Results

A. Dataset and Preprocessing

We evaluated our method on the Asu-Mayo Clinic Polyp Database [16] of *MICCAI 2015 Challenge on Polyp Detection* [1]. The dataset consists of videos with various frames and the videos are selected to display maximum variations in colonoscopy procedures (e.g., polyp variations, different resolutions, different detection strategies, existence of instruments information). The training dataset contains 20 colonoscopy videos with pixel-level annotated polyp masks in each frame. Among them, 10 videos have polyps inside and the other 10 videos have no polyp. There are totally 3799 frames with polyps. For the videos with polyps, each video contains a unique polyp. But this unique polyp disappears in most of frames and shows the maximum variations in different size, location, view and light. The testing dataset contains 18 videos with ground truth held out by the challenge organizers for independent evaluation.

[1]https://grand-challenge.org/site/polyp/

Due to the different resolutions of colonoscopy videos, we first resized all videos into fixed dimensions with spatial size of 570×320 before processing. We did not use padding in the 3D-FCN and the size of generated probability map was 31×15. Because probability values in the boundary of generated probability maps indicated the probabilities of the polyps in the corner of original colonoscopys, we did not do special processing of the pixels at the frame boundaries although a polyp may be located at the corner. To increase robustness and reduce overfitting, we utilized the strategy of data augmentation to enlarge the training dataset when training offline network. The augmentation operations, including rotation (rotating 90, 180, 270 degrees in the spatial plane) and translation (shifting the polyp locations by uniformly sampling values: $\Delta_s \sim (-10, 10)$ in the spatial plane and $\Delta_t \sim (-3, 3)$ in the temporal plane), were performed on extracted training sub-volumes. After data augmentation, we got about 85,000 positive training samples and we also extracted the same number negative training samples to train the *offline*-3D-Net.

B. Evaluation Metrics

We employed Precision (P) and Recall (R) to quantitatively evaluate the performance of our proposed polyp detection method. As low precision with high recall leads to heavy burdens for clinicians and low recall with high precision may result in late diagnosis of colon cancer, we also employed F1 score and F2 score to balance these two metrics. The above four metrics are defined as:

$$F1 = \frac{2PR}{P+R}, F2 = \frac{5PR}{4P+R},$$
$$P = \frac{N_{tp}}{N_{tp}+N_{fp}}, R = \frac{N_{tp}}{N_{tp}+N_{fn}}, \quad (4)$$

where N_{tp}, N_{fp} and N_{fn} denote the number of true positives (TP), false positives (FP) and false negatives (FN), respectively. Note that all the metrics are defined on polyp-level. A provided polyp detection is considered as a true positive if it falls inside the polyp masks; otherwise it is regarded as a false positive. A false negative is a polyp that has not been detected by the automated method.

C. Analysis of 3D-FCN

We investigated several different architectures of 3D-FCN to empirically identify a good architecture. Besides the proposed architecture above (Arch I in Table I), we also trained 3D-FCN with large convolution kernels and different receptive fields (Arch II in Table III and Arch III in Table IV). The Arch II and Arch III both employed large convolutional kernels with size of $5\times5\times5$ and Arch III had a different receptive field of $92\times92\times16\times3$. Table II shows the detection performance of offline model with different architectures. Note that all of these networks were trained from scratch for fair comparison. We can observe that the Arch I has better performance than Arch II and Arch III, which demonstrates that smaller convolution kernels are more efficient than larger convolution kernels.

TABLE II
THE DETECTION PERFORMANCE OF OFFLINE MODEL WITH DIFFERENT ARCHITECTURES.

Method	TP	FP	FN	Prec [%]	Rec [%]	F1 [%]	F2 [%]
Arch I	2289	1972	2024	53.7	53.0	53.4	53.2
Arch II	2203	3433	2110	39.1	51.1	44.3	48.1
Arch III	2005	1999	2308	50.1	46.5	48.2	47.1

TABLE V
RESULTS OF POLYP DETECTION ON ASU-MAYO DATASET USING THE PROPOSED FUSION MODEL AND THE OFFLINE NETWORK.

Method	TP	FP	FN	Prec [%]	Rec [%]	F1 [%]	F2 [%]
Offline-3D-Net	3053	835	1260	78.5	70.8	74.5	72.2
Fusion model	**3062**	**414**	**1251**	**88.1**	**71.0**	**78.6**	**73.9**

TABLE III
THE ARCHITECTURE OF DIFFERENT 3D-FCN (ARCH II).

Layer	Feature maps	Kernel size	Stride
Input	$102\times102\times16\times3$	-	-
Conv1a	$100\times100\times14\times64$	$3\times3\times3$	$1\times1\times1$
Pool1	$50\times50\times14\times64$	$2\times2\times1$	$2\times2\times1$
Conv2a	$48\times48\times12\times128$	$3\times3\times3$	$1\times1\times1$
Pool2	$24\times24\times12\times128$	$2\times2\times1$	$2\times2\times1$
Conv3a	$20\times20\times8\times256$	$5\times5\times5$	$1\times1\times1$
Pool3	$10\times10\times8\times256$	$2\times2\times1$	$2\times2\times1$
Conv4a	$6\times6\times4\times512$	$5\times5\times5$	$1\times1\times1$
Pool4	$3\times3\times2\times512$	$2\times2\times2$	$2\times2\times2$
Conv5	$1\times1\times1\times1024$	$3\times3\times2$	$1\times1\times1$
Conv6	$1\times1\times1\times2$	$1\times1\times1$	$1\times1\times1$

TABLE IV
THE ARCHITECTURE OF DIFFERENT 3D-FCN (ARCH III).

Layer	Feature maps	Kernel size	Stride
Input	$92\times92\times16\times3$	-	-
Conv1a	$88\times88\times14\times64$	$5\times5\times3$	$1\times1\times1$
Pool1	$44\times44\times14\times64$	$2\times2\times1$	$2\times2\times1$
Conv2a	$40\times40\times12\times128$	$5\times5\times3$	$1\times1\times1$
Pool2	$20\times20\times12\times128$	$2\times2\times1$	$2\times2\times1$
Conv3a	$16\times16\times8\times256$	$5\times5\times5$	$1\times1\times1$
Pool3	$8\times8\times8\times256$	$2\times2\times1$	$2\times2\times1$
Conv4a	$4\times4\times4\times512$	$5\times5\times5$	$1\times1\times1$
Pool4	$2\times2\times2\times512$	$2\times2\times2$	$2\times2\times2$
Conv5	$1\times1\times1\times1024$	$2\times2\times2$	$1\times1\times1$
Conv6	$1\times1\times1\times2$	$1\times1\times1$	$1\times1\times1$

D. Analysis of Offline and Online Learning

Fig. 5 shows some typical polyp detection results. In order to diagnose the role of the online representation learning, we show the detection results of both the fusion model and the offline network without integrating the online learning scheme. From the results shown in the first row of Fig. 5, we can see that both the proposed fusion model and the offline network can accurately single out polyps with variations in shape, color and texture from colonoscopy videos. The results highlight that the proposed 3D-FCN can tackle the large variations of polyps by exploring discriminative spatio-temporal feature representations. The second row presents some different detection results between the fusion model and the offline model. From these results, we can observe that the fusion model successfully removes some polyp-like false positives detected by the offline network. In addition, the fusion model integrating online and offline representation learning can even detect the polyps that

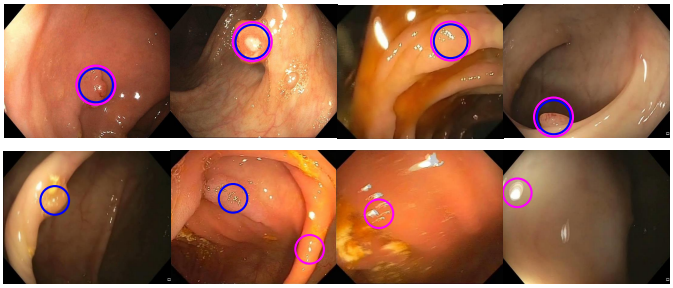


Fig. 5. Examples of polyp detection results. Blue and purple circles represent detection results of the fusion model and the offline network, respectively.

are neglected by the offline network. These results demonstrate the effectiveness of the online learning strategy aiming at dynamically learning and exploiting the specific features of the input video in order to improve the detection performance.

We further quantitatively analyze the detection performance of the offline network and the fusion model on the challenge dataset. The results are listed in Table V. It is observed that, our fusion model reduces around half number of false positives (414 vs. 835) compared to the offline model and hence significantly improves the precision (88.1% vs. 78.5%). The results corroborate that the online representation learning can efficiently reduce the polyp-like false positives generated by the offline network through leveraging the specific information extracted from the input testing video. In addition, the fusion model also detects more true positives than the *offline*-3D-Net and improves recall to some extent. Overall, the fusion model integrating offline and online representation learning achieves better performance on all four metrics than the offline network, which evidences the integration of the online and offline representation learning can greatly improve the detection performance. But due to the online model updating and the need of generating two probability maps, the fusion model has longer processing time than offline model.

E. Comparison with Other Methods

We compare the proposed polyp detection method with several other methods participating the challenge. The results are

TABLE VI
RESULTS OF POLYP DETECTION ON ASU-MAYO DATASET FROM DIFFERENT METHODS.

Method	TP	FP	FN	Prec [%]	Rec [%]	F1 [%]	F2 [%]
PLS	1594	10103	2719	13.6	36.9	19.9	27.5
CVC-CLINIC [11]	1578	3456	2735	31.3	36.6	33.8	35.4
OUS	2222	229	2091	90.6	51.5	65.7	56.4
ASU [16], [30]	2636	184	1677	**93.5**	61.1	73.9	65.7
CUMED	3081	769	1232	80.0	**71.4**	75.5	73.0
Fusion model (ours)	3062	414	1251	88.1	71.0	**78.6**	**73.9**

shown in Table VI[2]. The teams CVC-CLINIC [11] and PLS used hand-crafted features to locate polyps, while the teams CUMED and OUS employed 2D-CNN based approaches to automatically learn features from the training videos and then detected polyps. While the OUS team employed the traditional sliding windows strategy, the CUMED team adopted a segmentation-based strategy, where they first used a 2D-CNN to segment polyps in each frame and then located polyps based on the segmentation masks. The ASU team utilized a hybrid approach which integrates hand-crafted features and CNN based features [16], [30] . They first generated a set of polyp candidates using hand-crafted geometric features and then applied an ensemble of 2D-CNNs to classify each candidates.

We have three major observations from the results shown in Table VI. First, all the CNN based methods achieve better performance than the methods based on hand-crafted features, suggesting that the high-level features learned from CNN are more discriminative than the hand-crafted features. Second, the proposed method achieves the best performance on both F1 score and F2 score among all methods. The results further demonstrate the effectiveness of the proposed fusion strategy integrating online and offline representation learning in dealing with large variations of polyps and compensating the discrimination deficiency of offline models caused by limited specificity. Third, after carefully studying the results, we find that our proposed method has a better trade-off between precision and recall than 2D-CNN based methods. Our method achieves much higher recall performance, surpassing ASU and OUS by a large margin (about 10% and 20%) but our precision is lower than theirs. The higher precisions may be because the ASU team first used global geometric features to generate candidates with removing most of polyp-like false positives while OUS team set a high probability threshold for the final classification. However, as a trade-off, these schemes may increase the false negative and lead to a lower recall. On the other hand, our method outperforms the CUMED team by a large margin in term of precision (about 8%) while still achieving competitive recall performance (71.0% vs. 71.4%). Note that, in clinical practice, the balance between precision and recall of an automated detection approach is quite important. While low precision may increase doctors' workload for re-checking, low recall may cause mis-diagnosis or delay in diagnosis that prevents the early or timely treatment. This is why the challenge ranks the participants based on F1 score and F2 score. Overall, the challenge results demonstrate the discrimination capability of the proposed 3D-FCN and the effectiveness of the offline and online integration scheme in improving the detection performance.

We further compare the proposed methods with other methods on two subsets of the ASU-Mayo dataset: 1) a subset including videos with at least one frame containing polyp, and 2) a subset including videos with every frame containing polyp. While the first subset is composed of the most common cases in clinical practice, the videos in the second subset can be used to confirm the diagnosis and assist the subsequent interventions such as endometrial ablation. The results are shown in Table VII. The results of subset 1 are quite similar with the results reported in Table VI, where our method achieves the highest $F1$ and $F2$ score among all methods. As for the subset 2, we achieves 0 false positive and 100% precision, outperforming other methods by a large margin. This is attributed to that our method takes advantage of both spatial and temporal features extracted by the proposed 3D-FCN; the temporal features are quite important to detect polyps in a series of consecutive frames. The high precision on such a subset demonstrates the potential of the proposed method to be applied in computer-assisted interventions, where the proposed method can help detect and track the polyps for more precise operations. Moreover, our method can detect the polyp at least in one frame for all videos with polyps and thus has a relatively low miss-rate of individual polyp.

IV. DISCUSSION

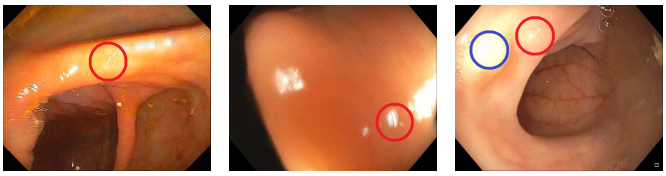
One of the main challenges for automated detection of polyps from colonoscopy videos lies in that there are a lot hard mimics in colonoscopy videos, such as bubbles, fecal content and specular spots. These hard mimics can seriously hinder the detection performance. A straightforward thought is to use some preprocessing methods to remove some of these hard mimics. However, these polyp mimics are very irregular and vary greatly in different colonoscopy videos; it is hard to use simple prepocessing methods to eliminate them. We have considered including them into the negative training samples to improve the performance of our method. However, this scheme needs us to manually annotate the positions of these mimics, which is out of the scope of this challenge because it needs extra labels. We therefore propose the online and offline representation learning integrated framework to reduce the influence of hard mimics through the online sample selection and online training. While the proposed integrated framework can achieve good results for discriminating hard

[2]The challenge result can be found in https://polyp.grand-challenge.org/results/

TABLE VII
RESULTS OF POLYP DETECTION ON SUB-DATASETS SELECTED FROM ASU-MAYO DATASET.

Videos with at least one frame containing polyp							
Method	TP	FP	FN	Prec [%]	Rec [%]	F1 [%]	F2 [%]
PLS	328	6953	2321	4.5	12.4	6.6	9.2
CVC-CLINIC [11]	195	1343	2454	12.7	7.4	9.3	8.0
OUS	651	55	1998	92.2	24.6	38.8	28.8
ASU [16], [30]	1218	92	1431	**92.9**	45.9	61.5	51.1
CUMED	1439	600	1210	70.6	**54.3**	61.4	57.0
Fusion model	1424	385	1225	78.7	53.8	**63.9**	**57.4**

Videos with every frame containing polyp							
Method	TP	FP	FN	Prec [%]	Rec [%]	F1 [%]	F2 [%]
PLS	1266	3150	398	28.7	76.1	41.6	57.2
CVC-CLINIC [11]	1383	272	281	83.6	83.1	83.3	83.2
OUS	1571	167	93	90.4	94.4	92.3	93.6
ASU [16], [30]	1418	40	246	97.2	85.2	90.8	87.4
CUMED	1642	149	22	91.7	**98.7**	95.0	97.2
Fusion model	1638	0	26	**100**	98.4	**99.2**	**98.7**



(a) (b) (c)

Fig. 6. Some failure cases of our framework. The red circles represent the detection results of our method while the blue circles represent the true polyps.

mimics, there are still some failure cases, especially when there are too many very similar mimics (e.g., colon walls) in the video, as shown in Fig. 6a. It is worth manually annotating these mimics and including them into the training samples to improve the performance. It is also observed that the image quality (e.g., image blur or overexposed regions) would influence the detection performance. Fig. 6b and Fig. 6c show the wrong detections due to the blurry image and overexposed regions. In the future, we shall investigate to utilize some image processing techniques (e.g., image deblurring and image normalization) to further improve the performance.

In recent years, deep convolutional neural networks (CNNs) have been widely applied in medical image analysis field and achieved remarkable success in many applications. We find that most of CNNs in medical image analysis field employ the architectures in natural image domain [20], [21] or follow the typical design principles employed in natural image processing applications (e.g., U-net in [25], DCAN in [26]). We think the reason is that these network architectures or design principles are summarized from rich design exploration and experiments; they are employed by many applications in the natural image domain and general enough to be extended to the medical image domain (i.e., colonoscopy videos). In our work, we also adopt some typical design principles (e.g, harnessing small convolution kernels, doubling the number of feature maps at downsampling step) in our network design. Our experimental results demonstrate the effectiveness of these design guidelines. It indicates that we can borrow the wisdom and successful experience in natural image domain for medical image analysis applications. Note that we do not employ the upsampling layers in our network. Instead, we use Eq. 2 to explicitly map the results back to the original locations in video clips. This is different from original fully convolutional networks for semantic segmentation tasks [37].

Training a deep 3D-FCN from scratch (i.e., the weights of networks are randomly initialized) is difficult because it requires a large amount of training samples. However, the insufficiency of training data is a well-known challenge of harnessing deep learning techniques in medical image analysis. Compared with millions of videos that can be acquired in natural video analysis tasks (for examples, $0.8M$ videos in YFCC100M dataset [44] and $1.1M$ sport videos in Sport-1M dataset [42] for detection and classification), we only have 20 training videos in this polyp detection challenge. It is difficult to solve this problem in many applications due to the high cost of data acquisition and labeling, not to mention that the small number of subjects for some rare diseases. In order to mitigate this problem, we used transfer learning (i.e., fine-tuning CNN models pre-trained from natural image dataset to medical image analysis tasks [20]). Most of studies [18], [20], [21] have demonstrated that transfer learning from the large scale annotated natural image datasets to medical image analysis applications has been consistently beneficial despite the difference between natural image dataset and medical image dataset. Therefore, we fine-tuned our 3D-FCN from a pre-trained model on Sport-1M. The big performance margin of fine-tuning model and randomly initialized model has demonstrated the effectiveness of fine-tuning.

In this work, we use the spatio-temporal features to automatically detect polyp in colonoscopy videos. Although it is not ready for the *in vivo* clinical use due to the processing time, our method can be further accelerated in the future. Specifically, we can investigate the following aspects for the acceleration: 1) using multi-process and multi-GPU techniques to process the different frames at the same time; 2) leveraging

some recently proposed model compression techniques, such as FitNet [45] and XNOR-Net [46], to reduce the computation time of each frame; 3) adjusting the online model training scheme to update the online model sample by sample, which can reduce the time of re-training in online model update. In addition, there are many other application scenarios for the proposed method besides the real-time *in vivo* polyp detection. For example, our method can provide alarm warnings to the operators in clinical practice. This alarm could remind doctors of coming back to re-identify the polyps. Our method can also be applied to offline processing of colonoscopy videos, which would help automatic document the operation process and efficiently construct a knowledge database for training new clinicians.

V. Conclusion

In this paper, we propose a novel online and offline 3D deep learning integration framework to automatically detect polyps from colonoscopy videos by leveraging 3D fully convolutional networks. The 3D networks can effectively learn spatio-temporal feature representations encoding more discrimination capability than features learned only from spatial information. More importantly, the fusion model integrating online and offline representation learning can significantly reduce the number of false positives and further improve the discrimination capability. Experiments on Asu-Mayo Clinic Polyp Database demonstrated the performance of our method and we achieved the best performance on F1 and F2 score metrics. The proposed fusion learning framework provides a new strategy to fill the gap between the large variation of testing data and the limited training data, which is a common challenge when employing supervised learning methods, especially deep neural networks, in data-driven medical image analysis tasks. Future investigations include evaluating our method on more clinical data and extending it to more detection tasks in endoscopic videos.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2016," *CA: A cancer journal for clinicians*, 2015.

[2] A. Leufkens, M. Van Oijen, F. Vleggaar, and P. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study." *Endoscopy*, vol. 44, no. 5, pp. 470–475, 2012.

[3] L. Rabeneck, H. B. El-Serag, J. A. Davila, and R. S. Sandler, "Outcomes of colorectal cancer in the united states: No change in survival (1986–1997)," *The American journal of gastroenterology*, vol. 98, no. 2, pp. 471–477, 2003.

[4] M. Ganz, X. Yang, and G. Slabaugh, "Automatic segmentation of polyps in colonoscopic narrow-band imaging data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 8, pp. 2144–2151, 2012.

[5] S. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. Karras, M. Tzivras *et al.*, "Computer-aided tumor detection in endoscopic video using color wavelet features," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 3, pp. 141–152, 2003.

[6] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A comparative study of texture features for the discrimination of gastric polyps in endoscopic video," in *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*. IEEE, 2005, pp. 575–580.

[7] L. A. Alexandre, N. Nobre, and J. Casteleiro, "Color and position versus texture features for endoscopic polyp detection," in *2008 International Conference on BioMedical Engineering and Informatics*, vol. 2. IEEE, 2008, pp. 38–42.

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

[9] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. C. De Groen, "Polyp detection in colonoscopy video using elliptical shape feature," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 2. IEEE, 2007, pp. II–465.

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

[11] 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," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.

[12] Y. Wang, W. Tavanapong, J. Wong, J. Oh, and P. C. de Groen, "Part-based multiderivative edge cross-sectional profiles for polyp detection in colonoscopy," *Biomedical and Health Informatics, IEEE Journal of*, vol. 18, no. 4, pp. 1379–1389, 2014.

[13] Y. Wang, W. Tavanapong, J. Wong, J. H. Oh, and P. C. De Groen, "Polyp-alert: Near real-time feedback during colonoscopy," *Computer methods and programs in biomedicine*, vol. 120, no. 3, pp. 164–179, 2015.

[14] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, A. Yousif *et al.*, "A colon video analysis framework for polyp detection," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1408–1418, 2012.

[15] 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," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, no. 2, pp. 283–293, 2014.

[16] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2016.

[17] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Sánchez, and B. van Ginneken, "Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1160–1169, 2016.

[18] H. Chen, Q. Dou, D. Ni, J.-Z. Cheng, J. Qin, S. Li, and P.-A. Heng, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 507–514.

[19] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

[20] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[21] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[22] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *MICCAI 2014*. Springer, 2014, pp. 520–527.

[23] H. Chen, Q. Dou, X. Wang, J. Qin, and P. A. Heng, "Mitosis detection in breast cancer histology images via deep cascaded networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[24] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 556–564.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[26] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: Deep contour-aware networks for accurate gland segmentation," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.

[27] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, and I. Išgum, "Automatic segmentation of mr brain images with

a convolutional neural network," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1252–1261, 2016.
[28] H. Chen, X. J. Qi, J. Z. Cheng, and P. A. Heng, "Deep contextual networks for neuronal structure segmentation," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
[29] D. Nie, L. Wang, Y. Gao, and D. Sken, "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2016, pp. 1342–1345.
[30] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 2015, pp. 79–83.
[31] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, and R. Tam, "Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1229–1239, 2016.
[32] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *arXiv preprint arXiv:1603.05959*, 2016.
[33] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P.-A. Heng, "Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1182–1195, 2016.
[34] R. Li, W. Zhang, H.-I. Suk, L. Wang, J. Li, D. Shen, and S. Ji, "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2014, pp. 305–312.
[35] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, 2013.
[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.
[37] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.
[38] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016.
[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[42] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.
[43] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.
[44] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," *arXiv preprint arXiv:1503.01817*, 2015.
[45] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
[46] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," *arXiv preprint arXiv:1603.05279*, 2016.

Lequan Yu (S'16) received the B. S. degree in Computer Science and Technology from Zhejiang University, Hangzhou, China, in 2015. He is currently a Ph.D. student in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. His reseach interests include medical image computing, deep learning and computer vision.

Hao Chen (S'14) received the B.S. degree in Information Engineering from Beihang University (BUAA) in 2013. He is currently a Ph.D. student in the Department of Computer Science and Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. His research interests include medical image analysis, deep learning, object detection and segmentation, etc.

Qi Dou (S'14) received the B. E. degree in Biomedical Engineering from Beihang University, Beijing, in 2014. She is currently a Ph.D. student in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong. Her research interests include medical image computing, deep learning, computer-aided detection, etc.

Jing Qin (M'16) is an assistant professor in School of Nursing, The Hong Kong Polytechnic University. He is also a key member in the Centre for Smart Health, SN, PolyU, HK. He received his Ph.D. degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2009. Dr. Qin's research interests include virtual/augmented reality for healthcare and medicine training, medical image processing, deep learning, visualization and human-computer interaction and health informatics. He has participated in more than 10 research projects and published more than 90 papers in major journals and conferences in these areas.

Pheng Ann Heng (M'92-SM'06) received the Ph.D. degree in computer science from Indiana University, Indianapolis, IN. He is currently a Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, where he is also the Director of the Virtual Reality, Visualization, and Imaging Research Centre. He is also the Director of the Research Center for Human-Computer Interaction, Shenzhen Institute of Advanced Integration Technology, Chinese Academy of Sciences, Shenzhen, China. His research interests include virtual reality applications in medicine, visualization, medical imaging, human-computer interfaces, rendering and modeling, interactive graphics, and animation.