Y-Net: A deep Convolutional Neural Network to Polyp Detection

Ahmed Kedir Mohammed[1]
mohammed.kedir@ntnu.no
Sule Yildirim-Yayilgan[2]
sule.yildirim@ntnu.no
Ivar Farup[1]
ivar.farup@ntnu.no
Marius Pedersen[1]
marius.pedersen@ntnu.no
Øistein Hovde [3]
Oistein.Hovde@sykehuset-innlandet.no

[1] Department of Computer Science
NTNU,
Gjøvik, Norway

[2] Department of Information Security and
Communication Technology
NTNU,
Gjøvik, Norway

[3] Department of Gastroenterology
University of Oslo,
Oslo, Norway

Abstract

Colorectal polyps are important precursors to colon cancer, the third most common cause of cancer mortality for both men and women. It is a disease where early detection is of crucial importance. Colonoscopy is commonly used for early detection of cancer and precancerous pathology. It is a demanding procedure requiring a significant amount of time from specialized physicians and nurses, in addition to a significant miss-rates of polyps by specialists. Automated polyp detection in colonoscopy videos has been demonstrated to be a promising way to handle this problem. However, polyps detection is a challenging problem due to the availability of limited amount of training data and large appearance variations of polyps. To handle this problem, we propose a novel deep learning method Y-Net that consists of two encoder networks with a decoder network. Our proposed Y-Net method relies on efficient use of pre-trained and un-trained models with novel sum-skip-concatenation operations. Each of the encoders are trained with encoder specific learning rate along the decoder. Compared with the previous methods employing hand-crafted features or 2-D/3-D convolutional neural network, our approach outperforms state-of-the-art methods for polyp detection with 7.3% F1-score and 13% recall improvement.

1 Introduction

Colorectal cancer is a major cause of morbidity and mortality throughout the world [9]. Early diagnosis is particularly relevant for colorectal cancer. The detection and removal of precancerous polyps in the colon may prevent later cancer. Screening the population for precursor lesions or manifest early colon cancer has been an important goal for decades [12][16]. Colonoscopy is often considered as gold standard because it allows an examination of the complete colon and it can remove pre-cancerous polyps immediately. It is a demanding procedure requiring significant amount of time from specialized physicians and nurses. Research works on colonoscopies have indicated significant miss-rates of 27% for small adenomas (< 5 mm) and 6% for adenomas of more than 10 mm diameter. There are multiple factors that contribute to missed polyps at colonoscopy including quality of bowel preparation and experience of the colonoscopist [6].

Colon polyps grow in two different shapes. Pedunculated polyps have mushroom-like tissue growths and sessile polyps have flat structures that sits on the surface of the mucous membrane. Automated polyp detection techniques [19][3][1] can be used to mitigate the miss-rates and compliment colonoscopist in fast diagnosis. However, high precision and recall automated polyp detection is a challenging problem since colon polyps exhibit different size, orientation, color, and texture. We depicted these variations in Fig. 1.

To detect colon polyps, researchers propose different methods [19][3] based on handcrafted features. These methods could not cope significantly with the problem of modeling distinctive features that differentiate polyps from normal mucosa. Therefore, deep convolutional neural networks (CNNs) are proposed to detect and localize colon polyps [4][20] that have presented very good performance. It is worth noting that CNN based methods require large labeled datasets for training. However, large medical datasets are not readily available due to proprietary rights and privacy concerns.

To cope with the aforementioned problem, we propose Y-Net, a novel encoder-decoder hybrid deep neural network that can be trained with limited amount of training data. Our proposed Y-Net deep model is capable of detecting and localizing different colon polyps regardless of orientation, shapes, texture, and size. In fact, our method is inspired by U-Net [14]. Our method is different from U-Net in that we use ensemble of encoders with and without pre-trained network and learn a decoder network and without extensive data augmentation. Y-Net model aims to learn decoder network from scratch while fine tuning encoder networks with different learning rates. Our method is conceptually simple, relying on the pretrained VGG network [17] as one of the encoders and a matched decoder network with the novel introduction of sum-skip-concatenation based connections to allow a much deeper network architecture for a more accurate segmentation. The key difference with the existing models is that we introduced a pre-trained encoder network that is augmented with untrained mirrored network and a decoder network that uses discriminative cost function to localize and detect polyps.

Our main contributions are: (1) A novel encoder-decoder network that uses a pretrained model for one of the encoders with mirrored untrained network and a decoder that is trained from scratch, making it more practical to train for polyp detection with limited amount of data. (2) A sum-skip-concatenation connection to a decoder network that allows re-use of pretrained encoder network weights. (3) Qualitative and quantitative experiments on MICCAI 2015 challenge on polyp detection [4][3][18] and comparison with state-of-the-art works [20], show that our method outperforms in polyp detection with 7.3% F1-score and 13% recall improvement.

2 Previous Work

There is a large literature dedicated to the topic of deep learning for image segmentation and detection. Our review here is brief due to space limitations, and is intended to highlight the broad approaches of existing polyp detection algorithms and to provide appropriate background for our work. For a current state-of-the-art and thorough review of deep learning for image segmentation and detection, please refer to [7]. Mainly for polyp detection, two

approaches are seen in literature: hand-crafted features with classifiers and deep-learning.
Hand-crafted features: CVC-CLINIC [3] proposed polyp localization algorithm by modeling the appearance of the polyps. Even though there are flat surface polyps but they assume all the polyps have protruding surfaces. Consequently, detection of the polyp is done through intensity valleys detection. They introduce a pre-processing step which helps to filter out other valley-rich structures like blood vessels. Similarly, Silva et al. [15] introduced a method which is influenced by the psycho-visual methodology of clinicians. Initially geometric features of polyps are exploited to get a region of interest (ROI) through the Hough transform. In the 2nd stage, texture features are used in an ad-hoc boosting-based classifier to filter out the ROI which does not contain any polyps. Generally, handcrafted features provide less correct detections and significantly more false alarms [4].
Deep learning approach: Different methods based on deep learning models are proposed [4][5][18][20]. Tajbakhsh et al. [18] came up with a two level approach for automatic polyp detection. Initially, geometric features like shape and size of polyps are exploited to get a set of candidate regions of having polyps. In the 2nd level, an ensemble of CNN are used for the classification. The output of the CNNs are averaged to get a probabilistic map for the existence of polyps in the frame. Yu et al. [20] proposed offline and online three-dimensional deep learning integration framework. With their approach they used online and offline 3-D representation learning to reduce the number of false positives and further improve the discrimination capability of the network for a specific video. Other teams on the MICCAI 2015 challenge [4] such as CUMED used VGG type architecture with down-sampling path and up-sampling path with auxiliary classifiers to alleviate the problem of vanishing gradients and encourage the backpropagation of gradient flow. Similarly, OUS used pretrained AlexNet model and CaffeNet, modifying the input size to 96 x 96 and applying sliding window on three scale of the original input image.

Our proposed Y-Net method is also based on deep learning model and is inspired by U-Net [14], a fully convolutional network. They used excessive data augmentation by applying elastic deformations to the available training images. However, our method explores multiple encoders or an ensemble of encoders with and without pre-trained network and learn a decoder network and without extensive data augmentation.

3 Methodology

Our overall approach is based on using robust pre-trained and un-trained encoder networks. The framework (illustrated in Fig. 2) consists of two fully convolution encoder networks which are connected to a single decoder network. The main goal for having two encoders network is to address the performance loss due to domain-shift from pre-trained network (natural images) to testing (polyp data), leading to degradation in performance. For example, a pre-trained models trained on natural images do not generalize well when applied to medical images. It is assumed that fine-tuning a pre-trained network works the best when the source and target tasks have high degree of similarity. Therefore, our approach focuses using the pre-trained model features optimally by slow fine-tuning the pre-trained network and aggressive learning on the second encoder for a better generalization on the test set. In the next sub-sections, we describe each of the network components, and then the losses used to train the network.

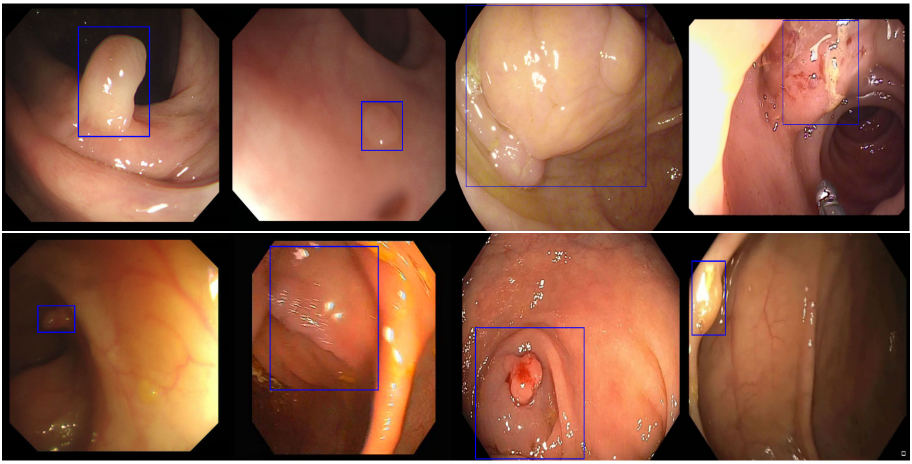

Figure 1: Appearance variability of polyps: Polyps vary in their shape, size and location within the large intestine. They may also have bleeding and abnormal patterns on surface.

3.1 Network Architecture

The architecture of our model is shown in Fig. 2. It consists of two contracting paths on the left, i.e. encoders, and expanding path to the right, i.e. a decoder, that matches the input dimension. The decoder outputs a binary mask segmentation of the polyp on the last layer.

Encoder one: It follows the typical architecture of VGG19 network [17], which has been widely used as the base network in many vision applications. This encoder uses a pre-trained weights of VGG19 trained on ImageNet dataset. The last fully connected layer of the network that was trained on 1000 ImageNet classes is truncated. Usage of a pre-trained model makes training easier and generalizes better in that, the pre-trained model already has learned features that are relevant to our own classification problems such as edges, curves and etc.

Encoder two: It follows the same VGG19 network architecture without the fully connected layer. The same input image is given to both encoders. It has 16 convolutional layers with 3×3 convolutions and 2×2 max-pooling layer. The weights of the network are initialized using Xavier normal initializer [8]. Moreover, it uses SELU activation function [11] to improve the back propagation as well as mitigates the vanishing and exploding gradients problem [13].

Decoder: The decoder network consists of five upsampling blocks and one final convolution block with a filter size of 1×1. Each upsampling block has the structure of Upsampling-concatenation followed by three blocks of CONV-SELU-BatchNorm(BN), except for the final layer which uses a 1×1 convolution with sigmoid for generating the final output mask. Compared with the other encoder-decoder architectures such as U-Net [14], our decoder is different in: (1) The decoder is not architecturally symmetric with the encoders. (2) The decoder is much deeper than the encoder. This design is due to the fact that with the limited polyp training data, a deeper decoder network would learn features from each scale of the encoder inputs that are concatenated with the same-scale decoder layer.

Sum-skip-concatenation: This can be generalized for any number of encoders as follows. Let $i \in N; 1 \leq i \leq d$ be the index for depth of an encoder network k and X is a convolution feature layer. Hence, $X_k^{i,j}$, represent the j_{th} index of convolution X at depth i. The values

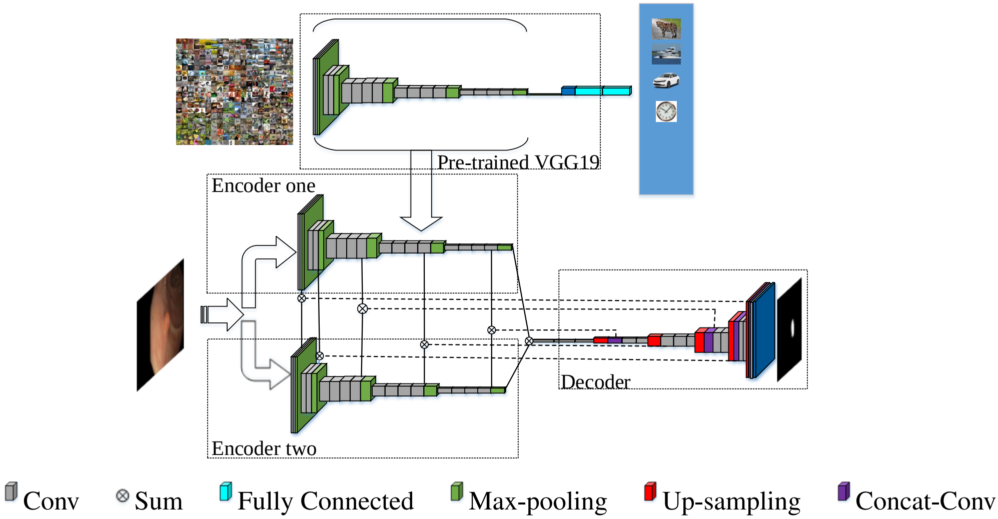

Figure 2: Y-Net: proposed architecture. The top row shows VGG19 network pre-trained on ImageNet dataset with 1000 classes such as cars, tiger and etc. The weights are transfered to encoder one network as shown above. Given an input image, it is fed to both encoders. The weights of last convolution at each depth of the encoder are summed and concatenated to the same spatial depth of the decoder. The output of the decoder is a binary label for foreground (polyp) and background (non-polyp).

of i and j depends on the specific architecture of the encoder network. Let $\mathcal{U}(\cdot)$ be up sampling operation. For VGG19 encoder network the j_{th} convolution in i_{th} depth of k_{th} encoder network $X_k^{i,j}$, the response of encoder convolution at depth X^i is estimated as follows:

$$X^i = \sum_{k=1}^{2} X_k^{i,j}, j = 2, i \leq 1 \leq 2, j = 4, i \leq 3 \leq 5 \tag{1}$$

For each depth of the decoder, the first feature input is computed by concatenating upsampled $\mathcal{U}(d-1)$ response of the decoder with the summed output of the encoder convolution feature X^i using (Eq. (1)). This connection has many advantages for training: (1) It can combine easily a pre-trained and un-trained network (Eq. (1)) with skip connection to a decoder. (2) by using different learning rate for both encoders, the un-trained encoder is trained aggressively without losing the learned pre-trained weights in training the decoder.

3.2 Model Learning and Implementation

Given the network architecture outlined above with one of the encoder pre-loaded with pre-trained VGG19 weights, we explain next the optimization objectives and training strategy.
Loss function: The output layer in the decoder consists of a single plane for foreground detected polyp. We applied convolution with sigmoid activation to form the loss. Let p and g be the set of predicted and ground truth binary labels, respectively. The weighted binary

cross-entropy and dice coefficient loss between two binary images is defined as:

$$\mathcal{L}(p,g) = -\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\lambda}{2}\cdot g_i \cdot \log p_i\right) + \left(1 - \frac{2\sum_{i=1}^{N}(g_i \cdot p_i) + \varepsilon}{\sum_{i=1}^{N}(p_i) + \sum_{i=1}^{N}(g_i) + \varepsilon}\right) \tag{2}$$

where λ and ε are false negatives (FN) penalty and smoothing factor, respectively. In order to penalize FN more than false positives (FP) in training our network for highly imbalanced data, the first term in Eq. (2) penalizes FN and the second term weighs FPs and FNs (precision and recall) equally. In other words, the second term is the same as the negative of F1-score. This is to avoid miss detection of polyps, as it is more critical to miss a polyp than giving a FP. Hence, the summed loss function gives a good balance between FN and FP.

Learning rates: Since the pre-trained encoder weights are initialized with VGG19, they are good, compared to randomly initialized weights, in extracting basic image features such as edges and curves. Therefore, it would be beneficial while training not to distort them too much. Hence, we propose encoder specific adaptive learning rates. The parameter update equation for RMSProp (Root Mean Square Propagation) gradient descent becomes

$$\theta_{t+1} = \theta_t - \frac{c \cdot \eta}{\sqrt{E[g^2] + \varepsilon}} \cdot g_t \tag{3}$$

where $c = 0.01$ for pre-trained encoder and $c = 1$ for encoder two and decoder, θ is a model parameter with a learning rate η and $E[g^2]$ is the running average of squared gradients. In this way, encoder two is learned aggressively while fine-tunning the pre-trained encoder.

4 Experiments

Dataset: Experiments are conducted on the ASU-Mayo clinic polyp database [19] of MICCAI 2015 Challenge on polyp detection. The dataset consists of 20 and 18 short segment colonoscopy videos for training and testing with pixel level annotated polyp masks in each frames respectively. The dataset contains colonscopy videos taken under variations in procedures (i.e. a careful colon examination while others show a fast colon inspection), has maximum variations in colonoscopy findings such as polyp variations, different resolutions, and existence of biopsy instruments. Among 20 training videos, ten videos have polyps inside and the other ten videos have no polyp. There are total 4278 frames with polyps in the training set and 4300 frames in test set.

Data augmentation: To increase robustness and reduce overfitting on our model, we increase the amount of training data. First, frames with polyp are doubled pre-training by applying random rotation (10° to 350°), zoom (1 to 1.3), translation in x,y (-10 to 10) and shear (-25 to 25) followed by centering the polyp and cropping padded regions. Second, during training, for each frame with polyp, random cropping of the non-polyp region as well as perspective transform is applied with probability of 0.3 and 0.4 with random horizontal and vertical flip respectively. We also tried applying contrast enhancement methods such as CLAHE (Contrast-limited adaptive histogram equalization) [21] and gamma correction, but that did not improve the detection accuracy.

Implementation Details: Our model is implemented on Tensorflow and Keras library with a single NVIDIA GeForce GTX 1080 GPU. Due to the different image sizes in the dataset, we first crop large boundary margins and resize all images into fixed dimensions with spatial size of 224×224 before feeding to both encoders and finally normalized to $[0,1]$. We use

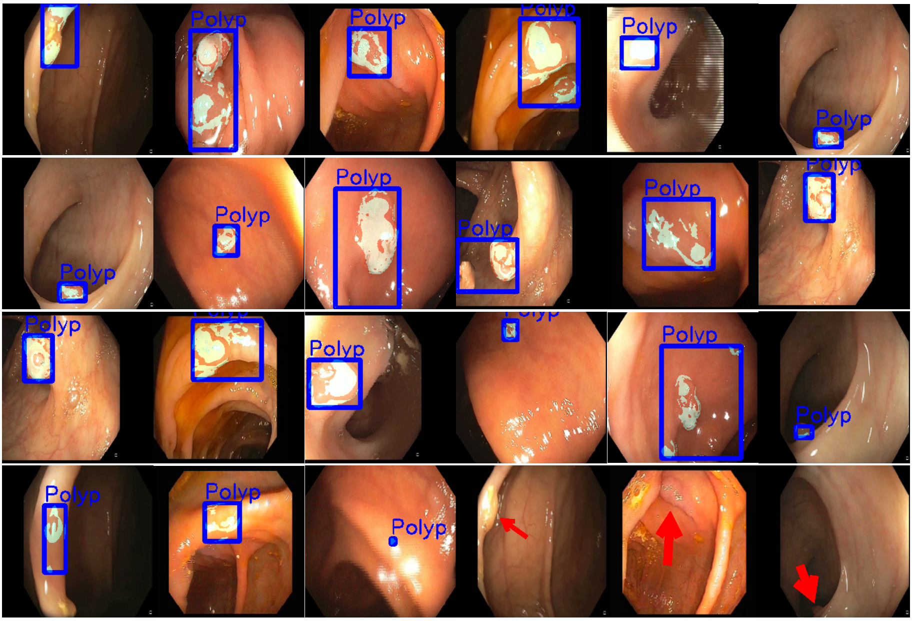

Figure 3: Sample detection result on ASU-MAYO test dataset. The images are chosen to show variations of the polyp appearance and performance of our approach. The last row shows failure cases. The first three images show FP and the last three show FN.

custom RMSProp (Eq. (3)) as the optimizer with batch size 3 and learning rate η set to 0.0001. We monitor dice coefficient and use *early-stop* criteria on the validation set error.

Evaluation Metrics: We evaluate the effectiveness of our model using recall and precision rate. If recall or true positive rate is low, then the model will miss finding polyps which can lead to late stage diagnosis for colorectal cancer. If precision is low, then it will add further examination and work for the gastroenterologist. Moreover, negative classes out number the positives with large margin, i.e. there are more frames without polyp than with polyp [2]. Hence, we employed F1-score and F2-score as they give a balance between missed polyps and false alarms. In order to calculate F1-score and F2-score we use the following equations:

$$F1 = \frac{2PR}{P+R}, F2 = \frac{5PR}{4P+R} \quad \text{and} \quad P = \frac{N_{TP}}{N_{TP}+N_{FP}}, R = \frac{N_{TP}}{N_{TP}+N_{FN}}, \tag{4}$$

where N_{TP} and N_{TN} are the number of true positives and negative and N_{FP} and N_{FN} are the number of false positive and negative, respectively. The detection is considered to be correct when the intersection over union (IoU) between detection box P and ground-truth bounding box G is greater than zero, otherwise the detection box is assumed as a false positive. Each pixel in the detection mask is considered as true detection if is above 0.9. IoU is formulated as:

$$IoU = \frac{P \cdot G}{\sum P + \sum G + \varepsilon} * 100\% \tag{5}$$

Table 1: Comparison of different polyp detection methods on ASU-MAYO dataset. The result for PLS and OUS are published in [4]. It can be seen that Y-Net has the lowest number of FN, and highest number of TP.

Method	TP	FP	FN	Prec[%]	Rec[%]	F1[%]	F2[%]
PLS	1594	10103	2719	13.6	36.9	19.9	27.5
CVC-CLINIC[3]	1578	3456	2735	31.3	36.6	33.8	35.4
OUS	2222	229	2091	90.6	51.5	65.7	56.4
ASU[19]	2636	184	1677	**93.5**	61.1	73.9	65.7
CUMED	3081	769	1232	80.0	71.4	75.5	73.0
Fusion[20]	3062	414	1251	88.1	71.0	78.6	73.9
Y-Net(Ours)	3582	513	662	87.4	**84.4**	**85.9**	**85.0**

4.1 Results

We compare the proposed method with several methods from MICCAI 2015 challenge on polyp detection [4] and with the recent work [20]. First, we show that the performance of our model is superior comparing to these methods considering F1-score, F2-score, and recall performance metrics. Then we present various approaches to cast insight on polyp detection and limitation of our method.

Comparison with the state-of-the-art: We compare Y-Net method with the state-of-the-art methods listed in Table 1. Polyp localize and spot team (PLS) on MICCAI 2015 challenge used a global hand-crafted image features followed by approximation of polyps with ellipses. CVC-CLINIC [3] directly models polyps with shape descriptors. Both PLS and CVC-CLINIC[3] use hand-crafted features followed by a classifier. ASU [19] uses both hand-crafted features for estimating edge map followed by ensemble of CNN for final polyp localization. The rest of the methods are based on deep learning approaches, where OUS uses pre-trained AlexNet and CUMED proposed custom CNN for segmentation. The last method we compared with, Fusion [20] uses 3D convolution with offline and online learning. Results in Table 1 demonstrate that our proposed method is superior to the other reference methods. The following observations can be made: (1) It is also interesting to observe that hand-crafted features have higher false positive rate than deep learning or hybrid approaches. We suspect this is because the polyps have varying shapes and appearance making it difficult for hand-crafting feature descriptor. (2) Our proposed method achieves the best performance in terms of F1-score, F2-score, and recall providing the maximum number of true detections and the minimum number false detections. (3) Our result suggests that hybrid fine tuning a pre-trained network and training from scratch a mirrored network gives a better performance when the domain of data are different such as in natural vs. medical images. Sample visual results are shown in Fig. 3.

Detection latency: One of the challenges in clinical colonoscopy when performing fast inspection is that some of the polyps can appear in a few frames and could be missed in the subsequent video. Therefore, it is also important to measure how quickly the first instance of the polyp can be detected by automated detection methods. Detection latency is the number of frames for our model to detect a polyp from the first appearance of a polyp in the scene, $\Delta T = t_2 - t_1$. The detection latency of our method for all test videos containing polyp is shown in Table 2. It is important to mention that for all videos, we measure the latency from

Table 2: Detection latency for videos containing only polyps on ASU-MAYO test dataset.

Detection Latency									
ΔT	8	0	0	18	0	0	0	0	0
Test video	Vid5	Vid6	Vid7	Vid8	Vid9	Vid10	Vid11	Vid12	Vid13

Table 3: Baseline comparison of pre-trained encoders: All the models are trained under the same data augmentation and common parameter settings

Method	Prec[%]	Rec[%]	F1[%]	F2[%]
U-Net(Trained from scratch, baseline)	90.8	39.2	54.7	44.2
U-Net(Pre-trained encoder VGG19, single encoder)	**96.2**	68.2	79.8	72.4
Y-Net(Ours)	87.4	**84.4**	**85.9**	**85.0**

first appearance of polyp on ground truth data.

4.2 Further insights and limitations

We further compare our method with pre-trained VGG19 single encoder with skip connection. In this case, the architecture is similar to U-Net except the encoder network of U-Net is replaced with a pre-trained VGG19 model. We trained the model together with the decoder. Moreover, we trained U-Net from scratch as baseline in order to analyze the effect of having a pre-trained network as an encoder. The result is summarized in Table 3. As it can be seen, using a pre-trained network as encoder part of U-Net improves the detection accuracy. Summing both pre-trained and un-trained encoder as proposed in this work results in the best F1 and F2-score for polyp detection. A unique characteristic of our method, compared with U-Net trained from scratch or with pre-trained encoder, is that it has a better compromise between recall and precision.

The last row of Fig. 3, shows some of false positives and false negatives of the proposed method. The sample frames are challenging in that there is a significant view point variation, shape, lighting and specular reflection. We have considered accounting for variation of light as well as contrast enhancement of the images as a preprocessing step. However, there is no significant difference in the accuracy.

Finally, it is important to note that, the proposed multiple(ensemble) of encoders can be generalized and extended to other architectures such as ResNet [10]. This can be done as discussed in section 3, by replacing the encoders with pre-trained and untrained ResNet network while keeping the decoder network shown in Fig. 2.

5 Conclusion

We address polyp detection problem in colonoscopy videos by proposing a new deep encoder-decoder approach. Our method relies on two-encoder networks that use a pre-trained VGG19 weights in the first encoder and random initialized weights in the latter. The two encoders are integrated with a novel sum-skip-concatenation operation. Moreover, we propose to use encoder based learning rate to efficiently use a pre-trained model. The approach is flexible in that, the pre-trained encoder used in this work can be replaced with other pre-trained models such as ResNet or similar. The proposed fusion of pre-trained and untrained encoder network with sum-skip connection to decoder 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 in medical image analysis tasks. Finally, experiment results on ASU-Mayo Clinic polyp database show that with the proposed multi-encoder framework, we achieved the best performance on F1 and F2 score metrics.

References

[1] Quentin Angermann, Aymeric Histace, and Olivier Romain. Active learning for real time detection of polyps in videocolonoscopy. *Procedia Computer Science*, 90:182–187, 2016.

[2] ChiChi Berhane and David Denning. Incidental finding of colorectal cancer in screening colonoscopy and its cost effectiveness. *The American Surgeon*, 75(8):699–704, 2009.

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized Medical Imaging and Graphics*, 43:99–111, 2015.

[4] Jorge Bernal, Nima Tajkbaksh, Francisco Javier Sánchez, Bogdan J Matuszewski, Hao Chen, Lequan Yu, Quentin Angermann, Olivier Romain, Bjørn Rustad, Ilangko Balasingham, et al. comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. *IEEE transactions on medical imaging*, 36(6):1231–1249, 2017.

[5] Mustain Billah, Sajjad Waheed, and Mohammad Motiur Rahman. An automatic gastrointestinal polyp detection system in video endoscopy using fusion of color wavelet and convolutional neural network features. *International journal of biomedical imaging*, 1, 2017.

[6] Stewart N Bonnington and Matthew D Rutter. Surveillance of colonic polyps: are we getting it right? *World journal of gastroenterology*, 22(6):1925, 2016.

[7] Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017.

[8] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[9] Fatima A Haggar and Robin P Boushey. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clinics in colon and rectal surgery*, 22(4):191, 2009.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 972–981, 2017.

[12] Reinier GS Meester, Chyke A Doubeni, Ann G Zauber, S Luuk Goede, Theodore R Levin, Douglas A Corley, Ahmedin Jemal, and Iris Lansdorp-Vogelaar. Public health impact of achieving 80% colorectal cancer screening rates in the united states by 2018. *Cancer*, 121(13):2281–2285, 2015.

[13] George Philipp, Dawn Song, and Jaime G. Carbonell. Gradients explode - deep networks are shallow - resnet explained. *CoRR*, abs/1712.05577, 2017. URL `http://arxiv.org/abs/1712.05577`.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[15] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. *International Journal of Computer Assisted Radiology and Surgery*, 9(2):283–293, 2014.

[16] Karen Simon. Colorectal cancer development and advances in screening. *Clinical interventions in aging*, 11:967, 2016.

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[18] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, pages 79–83. IEEE, 2015.

[19] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2016.

[20] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng Ann Heng. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. *IEEE journal of biomedical and health informatics*, 21(1):65–75, 2017.

[21] Karel Zuiderveld. Contrast limited adaptive histogram equalization. *Graphics gems*, pages 474–485, 1994.