Computers in Biology and Medicine 137 (2021) 104789

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A deep CNN model for anomaly detection and localization in wireless capsule endoscopy images

Samir Jain [a], Ayan Seal [a,*], Aparajita Ojha [a], Anis Yazidi [b,d,c], Jan Bures [e], Ilja Tacheci [e], Ondrej Krejcar [f,g]

[a] PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, 482005, India
[b] Department of Computer Science, OsloMet – Oslo Metropolitan University, Oslo, Norway
[c] Department of Plastic and Reconstructive Surgery, Oslo University Hospital, Oslo, Norway
[d] Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway
[e] Second Department of Internal Medicine-Gastroenterology, Charles University, Faculty of Medicine in Hradec Kralove and University Hospital Hradec Kralove, Sokolska 581, Hradec Kralove, 50005, Czech Republic
[f] Center for Basic and Applied Research, Faculty of Informatics and Management, University of Hradec Kralove, Hradecka 1249, Hradec Kralove, 50003, Czech Republic
[g] Malaysia Japan International Institute of Technology, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100, Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Deep convolutional neural network
Attention mechanism
Wireless capsule endoscopy
Anomaly detection
Localization

ABSTRACT

Wireless capsule endoscopy (WCE) is one of the most efficient methods for the examination of gastrointestinal tracts. Computer-aided intelligent diagnostic tools alleviate the challenges faced during manual inspection of long WCE videos. Several approaches have been proposed in the literature for the automatic detection and localization of anomalies in WCE images. Some of them focus on specific anomalies such as bleeding, polyp, lesion, etc. However, relatively fewer generic methods have been proposed to detect all those common anomalies simultaneously. In this paper, a deep convolutional neural network (CNN) based model 'WCENet' is proposed for anomaly detection and localization in WCE images. The model works in two phases. In the first phase, a simple and efficient attention-based CNN classifies an image into one of the four categories: polyp, vascular, inflammatory, or normal. If the image is classified in one of the abnormal categories, it is processed in the second phase for the anomaly localization. Fusion of Grad-CAM++ and a custom SegNet is used for anomalous region segmentation in the abnormal image. WCENet classifier attains accuracy and area under receiver operating characteristic of 98% and 99%. The WCENet segmentation model obtains a frequency weighted intersection over union of 81%, and an average dice score of 56% on the KID dataset. WCENet outperforms nine different state-of-the-art conventional machine learning and deep learning models on the KID dataset. The proposed model demonstrates potential for clinical applications.

1. Introduction

Computer-aided diagnostic (CAD) systems have seen tremendous growth in recent years. With the evolution of cyber-physical healthcare systems, the emphasis is on process automation in every field of diagnostics. Modern artificial intelligence techniques enriched with deep learning algorithms are major catalysts in these developments. Computer Vision-based diagnostic tools have shown remarkable improvements over the years in tasks like tumor detection in brain MRI [1], identification of gastro-intestinal malignancy [2], ulcers [3], polyps [4], tumors [5], bleeding [6] etc. WCE is a promising non-invasive painless method for the inspection of the digestive tract through captured videos [7]. In WCE, a tiny capsule equipped with a micro-sized camera (10 × 25 mm) is swallowed by the patient under examination. The video captured on the fly by the capsule is sent to a small receiver tied around the waist of the patient. WCE is a primary tool for the detection and diagnosis of anomalies like ulcers [8], bleeding regions [9], and polyps [10]. It enables examining those areas which conventional endoscopic procedures cannot reach. However, the downside of the WCE is the length of videos that may last from 8 to 10 h and may contain more than eighty thousand frames, with a frequency rate of 2 frames per second [11]. This makes the manual investigation of the entire tract a tedious and cognitively demanding task. The complete analysis of the video requires continuous concentration for very long hours which might

* Corresponding author.
E-mail address: ayan@iiitdmj.ac.in (A. Seal).

https://doi.org/10.1016/j.compbiomed.2021.104789
Received 11 April 2021; Received in revised form 18 August 2021; Accepted 18 August 2021
Available online 25 August 2021

overwhelm human attention. Due to this reason, the call for automation of the identification of abnormalities in video frames is of high demand. Recently, some product manufacturers have launched WCE systems with embedded software for identifying bleeding regions in video frames [12]. However, such systems struggle in identifying other types of anomalies. It is mainly because different types of anomalies pose challenges in identifying them using traditional image processing techniques due to the differences in color, texture, size and, shape.

Many Machine Learning (ML) techniques have been devised in the last decade for the detection of abnormal frames in WCE videos. The k-nearest neighbors (KNN), principal component analysis (PCA), support vector machines (SVM), random forest (RF), and artificial neural networks (ANN) are some of the most widely deployed classifiers for the identification of different types of anomalies. These ML algorithms are usually trained using manually extracted features such as color histogram [13], wavelet transform [14], local binary pattern (LBP) [15], Haralick texture features [16], scale-invariant feature transform (SIFT) [16], speed up robust features [17], and fractal dimensions (FD) [18] for the identification of WCE anomalies.

In recent years, deep learning has emerged as a promising tool for the detection of anomalous WCE images and localization of abnormal regions [4,11,19–21]. Notably, CNNs along with conventional feature extraction techniques have been used for anomaly detection such as lesion, hookworm, and bleeding [22,23]. Although deep learning techniques have shown improved performance in many cases, the problem we are tackling in this paper remains challenging due to the diversity of patterns and the complexity of textures of the abnormalities. The scarcity of labeled data and the computational requirements are also matters of concern when it comes to the applicability of these algorithms in WCE videos [24–26]. Until recently, most of the methods were developed to identify only a single type of anomaly. For automatic classification and localization of abnormal regions, a single method that can cover most of the anomalies in one shot is highly useful. Some of the recent approaches that address this problem include [11,27,28]. Iakovidis et al. [11] have suggested a CNN-based three-phase method that captures the features of a large number of anomalies in the WCE images. They have also estimated salient points in abnormal images which serve as guiding factors for the localization of abnormal regions. Gao et al. [29] have suggested a deep learning tool for the detection of outliers in the WCE image. They integrated a long short-term memory network with a CNN for learning the graphical patterns of ulcers, erythema, protruding lesions, polyp, and vascular malformation.

In this paper, a CAD model 'WCENet' is proposed to automatically identify and localize three different types of abnormalities in WCE images. The model works in two phases. In the first phase, a CNN model with an attention mechanism is trained to classify WCE images into 4 classes namely, inflammatory, polyp, vascular and normal. Images classified into abnormal classes are processed in the second phase for estimation of regions using a combination of a custom SegNet [30] and Grad-CAM++ [31]. The main contributions of the proposed work are summarized as follows:

- An attention-based CNN model is developed for the classification of WCE images into four categories namely, inflammatory, polyp, vascular (bleeding), and normal.
- A hybrid approach is used for the localization of anomalies in an image using both a custom SegNet [30] and Grad-CAM++ [31]. The hybrid approach provides higher precision than the individual methods.
- The proposed model demonstrates superior performance in terms of classification and localization performance in comparison with other state-of-the-art methods.

The rest of the paper is organized as follows. Section II gives a brief overview of related work. Section III is devoted to the presentation of the proposed model for anomaly detection and localization in WCE images. In Section IV, experimental results are presented and compared with some state-of-the-art techniques. Section V concludes the paper and delineates some future work.

2. Related work

Existing research in the field of WCE covers detection of anomalies like bleeding [22,32], polyps [33–35], hook-worms [19], ulcers [8], tumors [36] etc. Pixel-based approaches have been found effective for bleeding detection since color plays an essential role in this type of anomaly (see e.g. Refs. [12,32,37,38]). Some of the researchers have used the conversion of RGB images to a different color space to improve classification accuracy. Shah et al. in Ref. [39] have opted for HSI color space over RGB for the classification of bleeding and non-bleeding images using a color threshold (see also [6,12]). Shape and texture analysis proves to be very useful in the characterization of polyps, ulcers, inflammatory or vascular disorders, and tumors. Bag of visual words [32], color histogram [38], LBP [40], wavelet transform [36], FD [18] are some of the approaches relying on these types of features. Most of the image classification techniques mentioned above are either threshold-based where the count and the intensity of pixels are used [12] or use ML algorithms like KNN [32,37], SVM [13,17], and ANN [14,41]. Cong et al. [42]. have suggested a new variant of SVM termed as Deep Sparse SVM (DSSVM) for the identification of abnormal frames. DSSVM is trained on color and texture features that were extracted from image super-pixels. A group sparsity criterion is defined for the removal of irrelevant features by assigning a higher weight to important features. This model is found to be quite useful when the computed features exhibit redundancies.

Deep learning methods have recently emerged as powerful solutions in a variety of applications including WCE video analytics (see, for example [9,25,31,39,43]). Sekuboyina et al. [27]. have used the CIE-Lab color space representation of WCE images to train a CNN classifier on the patches falling in the normal and abnormal regions of WCE images. Once the CNN is trained, it locates the abnormal regions in an image by identifying and combining abnormal patches. Other methods that utilize CNN include [4,11,22,28]. CNN-based autoencoders have also been popular alternatives for the extraction of features from WCE images [10]. A stacked sparse autoencoder (SSAE) has been presented by Yuan et al. [10] for polyp detection. The accuracy of the model is enhanced using an image manifold constraint (SSAEIM) with the idea that the images belonging to the same class share common features whereas the images from different classes exhibit high variances in the learned features. Banik et al. [4] have suggested an integrated method for the segmentation of polyps by deploying a CNN and a level set method. A modified level set method termed as Local Gradient Weighting embedded Level Set Method (LGWe-LSM) is introduced for surface and shape analysis which suppresses high-intensity regions that may procreate false positives. A 16-layer deep CNN for the detection of polyps in colonoscopy images has been projected by Rahim et al. [33]. They have introduced generalized intersection over union to deal with scale invariance issues.

Deep neural networks are data-hungry and unfortunately, most of the publicly available WCE datasets are of a relatively small size as compared to the general category datasets on which deep learning models have shown remarkable performance. To overcome the problem of small datasets, the transfer learning approach has been used by many authors. Li et al. [44] have exploited the effectiveness of transfer learning for gastrointestinal bleeding detection on a small-size imbalanced endoscopy image dataset. Riberio et al. in Ref. [45] have considered pre-trained models AlexNet, VGGNet, and GoogLeNet on ImageNet and Pascal VOC datasets for the classification of colonic polyps. Shin et al. [46] have also utilized the transfer learning approach for the localization of polyps using Faster RCNN with Inception-Resnet (InceptionV4) [47]. In the method devised by Sadasivan et al. [28], a patch-based CNN is suggested where patches from normal and abnormal

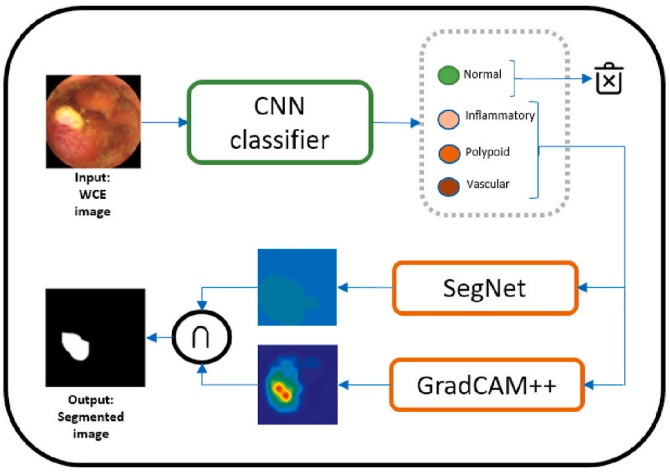

Fig. 1. Schematic block diagram of proposed WCENet.

regions of WCE images are extracted and used for the localization of anomaly. The CNN is trained to classify the patches as normal or abnormal. Ghosh et al. [43] presented a two-stage system for bleeding detection and localization in WCE images. In the first stage, a CNN with AlexNet architecture is used to classify bleeding and non-bleeding (normal) frames. In the next stage, bleeding regions are segmented using VGG16 based SegNet [30]. AlexNet is an old architecture and several efficient CNN models have been tendered over the years that show significant improvement over AlexNet [48]. Further, in recent years, some CNN models with attention mechanisms have also been explored in different domains which were shown to help improve the prediction performance of the models.

A review of the existing literature indicates that a handful number of generic methods have been presented to deal with multiple types of anomalies simultaneously. Further, deep CNN models used in the classification of anomalies have large memory footprints and demand high computational resources. In this paper, an attention-based CNN is proposed which detects four different types of anomalies in WCE images. To locate the abnormal regions in identified images, a hybrid of two standard approaches is applied, namely, Grad-CAM++ [31] and SegNet [30]. The proposed method provides improved classification accuracy and localization results with higher precision levels compared to legacy methods.

3. WCENet: an anomaly detection and localization model

In this section, an automatic anomaly detection and localization model 'WCENet' is proposed for WCE images. Fig. 1 presents the schematic diagram of the model that consists of (i) a base CNN model using an attention-based mechanism that classifies the images into four categories (ii) a custom SegNet [30] model for semantic segmentation of abnormal regions in images (iii) a supporting mechanism to the segmentation model using Grad-CAM++ [31] applied on the trained classification model.

The 11-layer attention-based CNN model is trained on a WCE image dataset to classify images into four categories namely, inflammatory, polyp, vascular, and normal. The attention mechanism focuses on dominant features and suppresses the insignificant ones. All images that are classified as abnormal by the CNN model are then passed onto the localization framework which consists of two parallel mechanisms (i) a SegNet based [30] CNN model that is trained to segment the pixels falling in the abnormal region of an image, and (ii) analysis of the class activation maps of the images using Grad-CAM++ [31] applied on the trained WCENet. The set of pixels that are identified by both methods as abnormal are finally selected for the localization of the anomaly in the image. Details of SegNet and Grad-CAM++ are given in Section 3.2.

3.1. The classification network

The classification model of WCENet is a CNN with an attention mechanism having 7 convolution (Conv) blocks and 4 fully connected (FC) blocks as shown in Fig. 2. A Conv block consists of a Conv layer followed by a batch normalization + ReLU layer that fends off the model from overfitting. The Conv blocks are grouped into 4 modules. The first module consists of two Conv blocks, with 32 filters, each of size 3×3 in both the Conv layers. Max-pooling is applied at the output of the second Conv block and then an attention block is added. The second and the third modules have an identical structure except for the number of filters. There are 64 filters in each of the Conv layers in the second module and 128 filters in the Conv layers of the third module. The fourth module has only one Conv block followed by max-pooling. Since the anomalies in the WCE images can be present at the boundaries also, the role of boundary pixels is crucial. Therefore, 'SAME' padding is used in the convolution operation throughout the Conv layers. As depicted in Fig. 2, an attention block is added after each of the first three Conv modules. The attention mechanism is induced from the concept of a convolution block attention module (CBAM) [49] which is briefly discussed in Section 3.1.1, the output of the fourth module goes to the global average pooling (GAP) layer. GAP layer contributes to limit the computation of

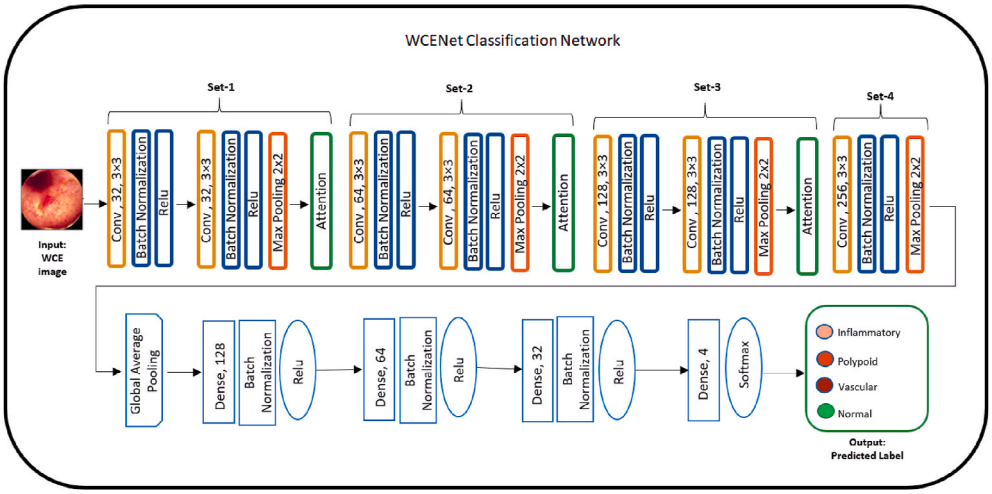

Fig. 2. Schematic block diagram of WCENet classifier.

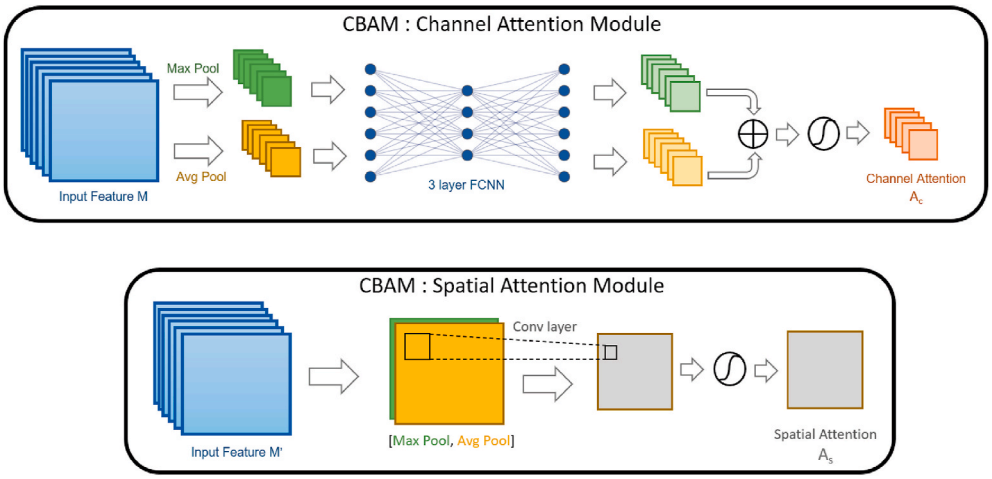

Fig. 3. Schematic block diagram of CBAM attention sub-modules.

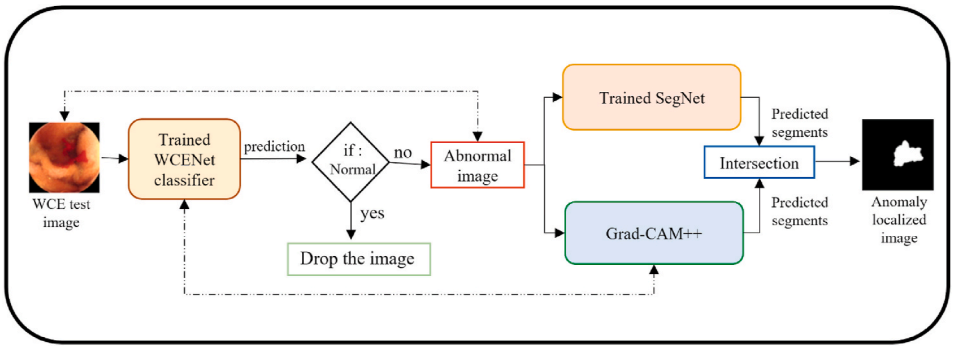

Fig. 4. A flow diagram of localization framework in WCENet.

parameters as compared to the flatten layer to make the network learn faster. Then four FC layers are added to the tail of the network with 128, 64, 32, and 4 units, the last layer is the (output) classification layer. The first three FC layers are coupled with batch normalization and ReLU activation whereas the output layer uses softmax activation. Input to the WCENet is a 3-channel, 24-bit RGB image of size 320 × 320.

3.1.1. Convolution block attention module

The CBAM [49] is a lightweight and generalized attention module that can be integrated easily with any CNN. CBAM derives attention maps from channels and spatial dimensions of intermediate feature maps. To understand how the attention mechanism works, consider an intermediate feature map $M \in \mathbb{R}^{C\times I\times J}$, where C denotes the number of channels in M and $I \times J$ is the feature dimension. CBAM extracts 1-D channel attention map $A_c \in \mathbb{R}^{C\times 1\times 1}$ and a 2-D attention map $A_s \in \mathbb{R}^{1\times I\times J}$ as shown in Fig. 3. The final feature map with attention is computed with the following series of operations:

$$M^{'} = A_c(M) \otimes M, \tag{1}$$

$$M'' = A_s(M^{'}) \otimes M^{'}, \tag{2}$$

where the symbol $\otimes$ denotes the element-wise multiplication of two matrices. Here, M'' is the final feature map with the attention that can be considered as the refined output. The channel attention module targets 'which' information inside the image is meaningful whereas the spatial attention module focuses on 'where' the meaningful information is located in the image. To compute the channel attention, the spatial dimensions are squeezed by aggregating the spatial information using max-pooling and average-pooling. These two spatial context feature descriptors M^c_{avg} and M^c_{max} are then forwarded to a single hidden layer fully connected NN (FC-NN) to produce $A_c \in \mathbb{R}^{C\times 1\times 1}$. The hidden activation can be fixed to a size of $\mathbb{R}^{C/n\times 1\times 1}$ where the size can be reduced by n. Each descriptor is passed through the shared FC-NN and the outputs of both the descriptors are added element-wise to get A_c which is then multiplied to each element of M along the channel axis to get M' as given in Eq. (1). Next, the spatial attention is computed on M' by max-pooling and average-pooling feature maps at each pixel location across all the channels producing M^s_{avg} and $M^s_{max} \in \mathbb{R}^{1\times I\times J}$. Convolution operation with filters of size f is then performed on the two-channel feature descriptor formed after concatenation of M^s_{avg} and M^s_{max} to produce $A_s \in \mathbb{R}^{1\times I\times J}$. Finally, element-wise multiplication of A_s is performed on each channel C of $M^{'} \in \mathbb{R}^{C\times I\times J}$ to produce a refined feature map M'' as given in Eq. (2). In this way, a CBAM can be easily applied to the output of any intermediate layer of a CNN.

3.2. Anomaly localization framework

In the second stage of WCENet, an abnormal image (labeled as abnormal after classification) is analyzed further for the localization of anomalies. The localization is achieved through a hybrid approach using two standard localization methods namely SegNet [30] and Grad-CAM++ [31]. SegNet is a deep encoder-decoder architecture that provides pixel-wise segmentation of an image. In the present paper, a custom CNN is used in the encoder and the decoder parts of SegNet. The second method is Grad-CAM++ which uses the trained WCENet classifier network and generates the class activation maps that highlight the salient region in an input image. The outputs of both the methods are then fused to get the segmented region in the WCE image for the

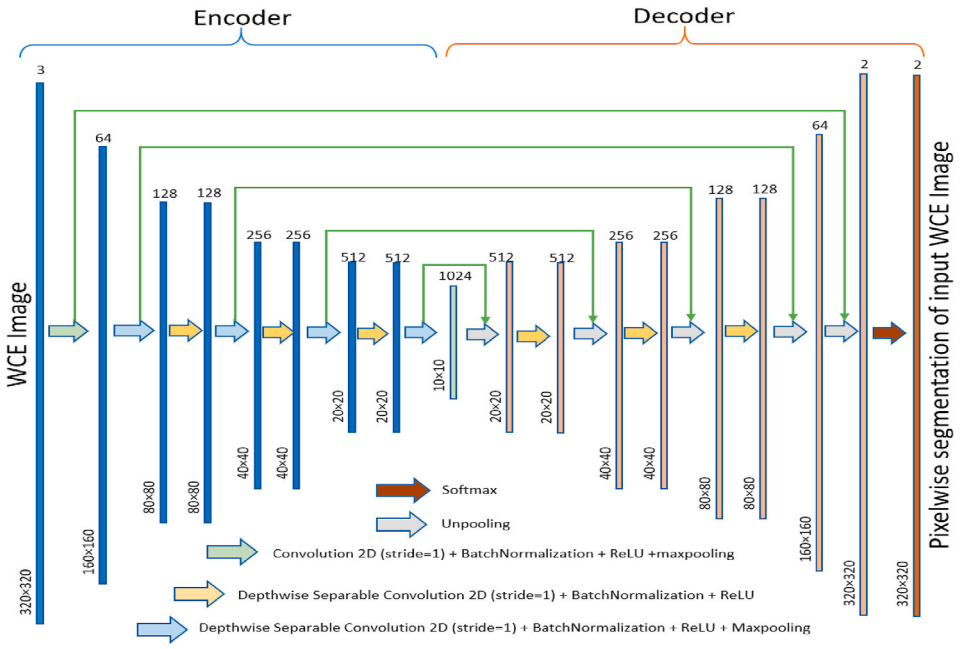

Fig. 5. SegNet architecture for localization of anomaly in WCE images.

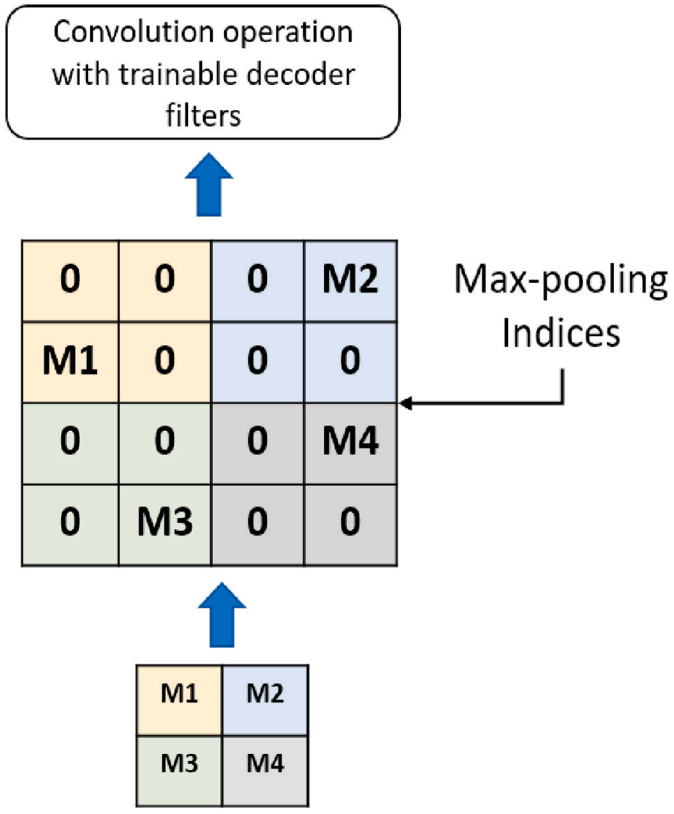

Fig. 6. A SegNet decoder employing max-pooling indices during unpooling.

localization of the anomaly in an image. The localization framework is presented in Fig. 4.

3.2.1. SegNet architecture

Gastrointestinal (GI) anomalies differ in their visual appearances. They exhibit irregular shapes, different colors, and textures which make them difficult to locate in an image by computer vision techniques. Pixel level segmentation can be useful for identifying such anomalous regions. In this paper, a customized CNN-based segmentation model inspired by the idea of SegNet [30] is proposed. The segmentation model trained on a WCE dataset can differentiate the pixels belonging to an abnormal or a normal region in an image. As mentioned earlier, the SegNet architecture consists of an encoder and decoder network which can be a CNN or FC-NN followed by a final pixel-wise classification layer.

SegNet model developed in the proposed work is shown in Fig. 5. In this network, depthwise separable convolutions (DSC) are used for the extraction of features. DSC makes the model lightweight by drastically reducing computations and model size. Both the encoder and the decoder are composed of 8 depthwise separable convolutions (DSC) layers. Each DSC layer is followed by the batch normalization layer and a ReLU activation layer. The DSC layer is configured with 3×3 filters and a stride $= 1$. Further, a 2×2 max-pooling layer with stride $= 2$ is used in the network. All the DSC layers are followed by batch normalization and ReLU layers. The decoder is a mirror network configured with upsampling layers in place of max-pooling layers. Here, the upsampling is performed as in the original SegNet. Since the max-pooling operation in the encoder leads to the loss of the spatial resolution, SegNet stores the location of the maximum value for each feature map during the max-pooling operation as shown in Fig. 6. The green lines connecting the encoder-decoder layers in Fig. 5 show max-pooling indices shared by the encoder and decoder which are used during the upsampling operation. The decoder generates sparse-feature maps which are then convolved with the filter banks to densify them. The decoded output is passed through the softmax layer that generates a $K-$ channel image of the probability values at pixel locations where K is the number of classes. Finally, a segmented image is obtained where each segment corresponds to an anomaly class giving the maximum probability at each pixel.

3.2.2. Grad-CAM++

In contrast to SegNet, which is trained explicitly on the datasets with ground-truths for pixel segmentation in an input image, Grad-CAM++ [31] works on a trained classifier, and builds a heatmap to identify the region of interest in an input image. Grad-CAM++ performs an analysis of the class activation maps (CAMs). Zhou et al. [50] suggested that various layers of a CNN can be exploited as object detectors using CAM if the GAP layer is utilized with the combination of weighted feature maps produced just before the softmax classification layer (penultimate

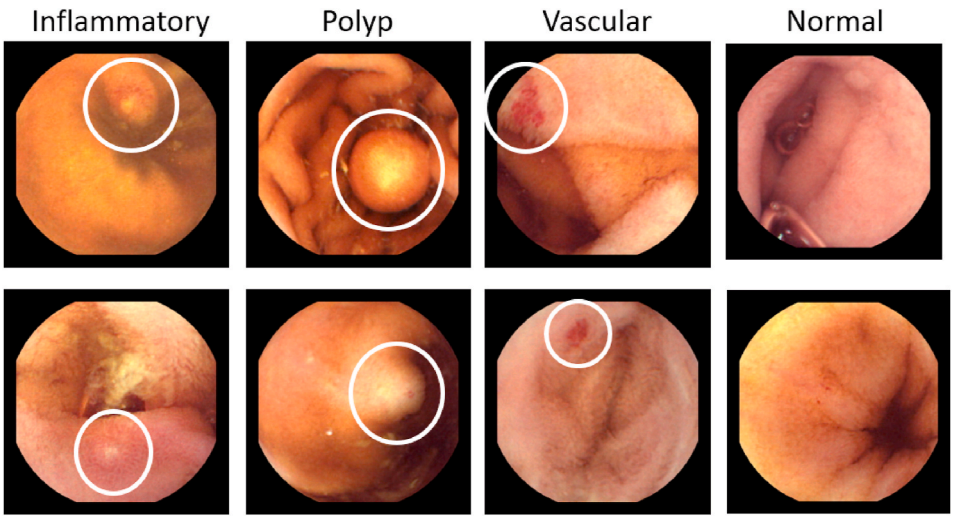

Fig. 7. Some samples of images categorized in four classes where anomalies present in the abnormal class are located with the white circle.

Table 1
Description of the KID dataset (KID-I and KID-II combined).

Class	# Images in the dataset	# Images after augmentation
Inflammatory	241	1266
Polyps	50	1293
Vascular	350	1243
Normal	728	1300

Table 2
Performance of WCENet with and without attention mechanism.

WCENet	Accuracy	Precision	Recall	F1-score
Without attention	0.97	0.96	0.97	0.97
With attention	0.98	0.98	0.98	0.98

layer). This method allows generating a heatmap that highlights the pixels in the image that have participated in assigning a particular class to the image.

Let the penultimate layer produce L feature maps, $M^l \in \mathbb{R}^{u \times v}$ of size $u \times v$. To apply the spatial GAP on these feature maps, the linear combination given in Eq. (3) is used to compute a score S^c for each class c.

$$S^c = \sum_l \overbrace{w_l^c}^{\text{class weights}} \underbrace{\frac{1}{Q}\sum_i \sum_j \overbrace{M_{ij}^l}^{\text{feature map}}}_{\text{GAP}}, \tag{3}$$

where Q is the number of pixels in the activation map M. The class-specific localization map L^c at each location (i, j) is calculated using Eq. (4).

Table 4
Parametric configuration of the ML classification methods compared with WCENet classifier.

Author	Method	Feature Length	ML classifier	Hyperparameters
Yuan et al. [40]	SIFT + CLBP	120	SVM	Cubic SVM kernel
Ghosh et al. [38]	CHOBS	4096	KNN	K = 1
Jain et al. [18]	DBC (FD)	2025	RF	# Estimators = 500

Table 5
Classification performance of nine state-of-the-art techniques and WCENet with 5-fold cross validation.

Method	Accuracy	Precision	Recall	f1-score	AUC
Sekuboyina et al. [27]	0.53	0.51	0.98	0.67	0.52
Georgakopoulos et al.(patch CNN) [55]	0.58	0.55	0.96	0.70	0.57
Sadasivan et al. [28]	0.57	0.54	0.99	0.71	0.59
Jia et al. [9]	0.90	0.92	0.85	0.87	0.85
Iakovidis et al. [11].	0.92	0.95	0.91	0.93	0.94
Ghosh et al. [38]	0.72	0.70	0.73	0.72	0.74
Yuan et al. [40]	0.82	0.79	0.85	0.82	0.83
Jain et al. [18]	0.83	0.85	0.84	0.84	0.84
Ghosh et al. [43]	0.94	0.96	0.94	0.94	0.95
WCENet	0.98	0.98	0.98	0.98	0.99

Table 3
Parametric configuration of six state-of-the-art CNN classification models and the proposed WCENet classifier.

Author	# CNN Layers	# Pooling Layers	# FC Layers	Hyperparameters (opt, lr, bs, ep)	# Trainable Parameters
Sekuboyina et al. [27]	3	2	3	Adam, 0.01, 100, 100	7,032
Georgakopolous et al. [55]	5	2	3	SGD, 0.001, 100, 100	115,486
Sadasivan et al. [28]	3	2	3	Adam, 0.001, 64, 200	7,032
Jia et al. [9]	3	3	2	SGD, 0.01, 100, 200	10,697,060
Iakovidis et al. [11]	5	4	3	SGD, 0.001, 50, 200	286,344
Ghosh et al. [43]	5	3	3	SGD, 0.01, 16, 200	528,078,924
Proposed WCENet	7	4	4	SGD, 0.001, 16, 200	656,010

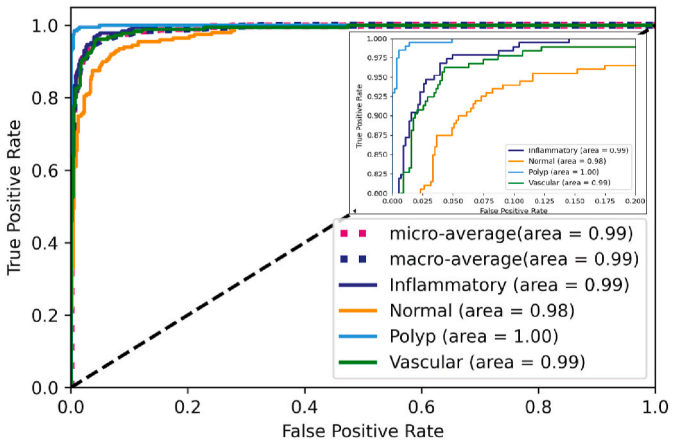

Fig. 8. ROC curves of WCENet.

Table 6
Performance comparison of segmentation models with different base architectures.

Method	Encoder Architecture	$F_w IoU$	MIou	Dc
SegNet	VGG16	0.79	0.60	0.48
SegNet	ResNet50	0.80	0.60	0.51
SegNet	MobileNetV1	0.81	0.61	0.55
SegNet	Custom CNN	0.81	0.60	0.55
UNet	VGG16	0.77	0.50	0.45
UNet	ResNet50	0.78	0.53	0.46
UNet	MobileNetv1	0.79	0.56	0.48
UNet	Custom CNN	0.79	0.56	0.48
PSPNet	VGG16	0.74	0.44	0.39
PSPNet	ResNet50	0.77	0.51	0.45
PSPNet	MobileNetv1	0.79	0.49	0.47
PSPNet	Custom CNN	0.78	0.49	0.48

$$L^c_{i,j} = \sum_l w^c_l \cdot M^l_{ij} \tag{4}$$

Note that the CAMs help visualizing the outputs of the last convolution layer only, which is a limitation. Also, the process involves training the linear classifier for each class. These issues were addressed in a method known as Grad-CAM introduced by Selvaraju et al. [51]. Instead of training multiple classifiers for the class-specific weight computation, the weights w^c_l for a specific activation map M^l and the class c are computed using Eq. (5).

$$w^c_l = \underbrace{Q \cdot \frac{\partial S^c}{\partial M^l_{ij}}}_{\text{gradients}} \qquad \forall i,j \tag{5}$$

The above formulation computes the weights w^c_l independent of the spatial locations (i, j) of a specific activation map M^l. The limitation of CAM is overcome by taking the GAP of the partial derivatives ∂M^l_{ij}. Therefore, the weights w^c_l are computed using Eq. (6).

$$w^c_l = \frac{1}{Q} \sum_i \sum_j \frac{\partial S^c}{\partial M^l_{ij}} \tag{6}$$

Grad-CAM's performance deteriorates when there are multiple instances of the same object in an image. This is quite common in WCE images where anomalies can be present at different locations in the same image. Further, the localization through Grad-CAM is observed to miss some portion of the region of interest, probably due to the unweighted average of partial derivatives. Grad-CAM++ introduced by Chattopadhyay et al. [31] overcomes this issue by modifying Eq. (6) as follows.

$$w^c_l = \sum_i \sum_j \underbrace{\beta^{lc}_{ij}}_{\text{gradient weights}} ReLU\left(\frac{\partial S^c}{\partial M^l_{ij}}\right) \tag{7}$$

The main idea behind the above formula is that w^c_l apprehends the significance of a particular activation map M^l. For a given activation map M^l, the positive gradient at location (i, j) makes the class score S stronger with increasing pixel values. Therefore, the linear combination of partial derivatives over each pixel in an activation map M^l will show the relevance of that map for class C. The formulation in Eq. (6), computes the weighted average of gradients w^c_l in contrast to the GAP performed using Eq. (7). The class score S^c is computed using Eq. (8) which

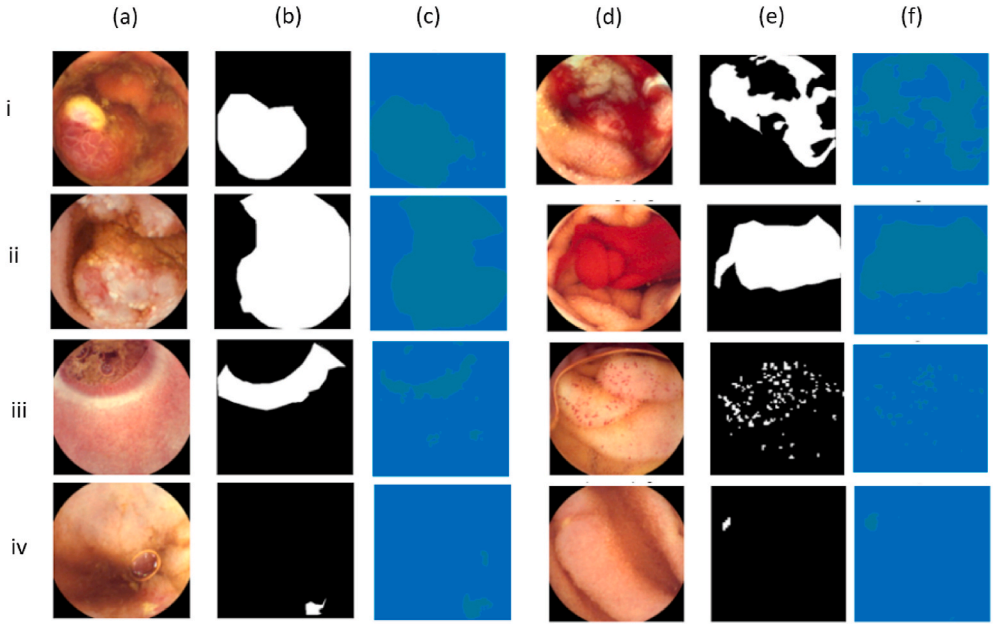

Fig. 9. (a,d) Original WCE image with anomalies. (b,e) Corresponding ground truths of WCE images. (c,f) Segmentation results produced by SegNet model.

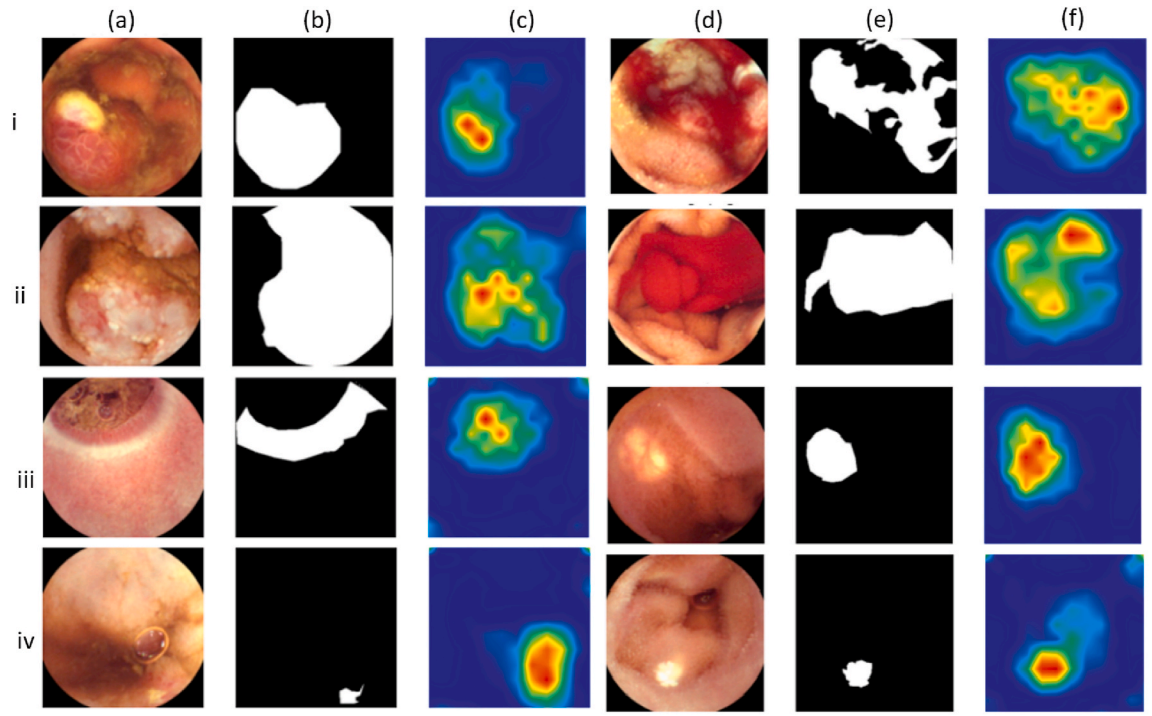

Fig. 10. (a,d) Original WCE image with anomalies. (b,e) Corresponding ground truths of WCE images. (c,f) Heatmaps through GradCAM++ applied on trained WCENet classifier.

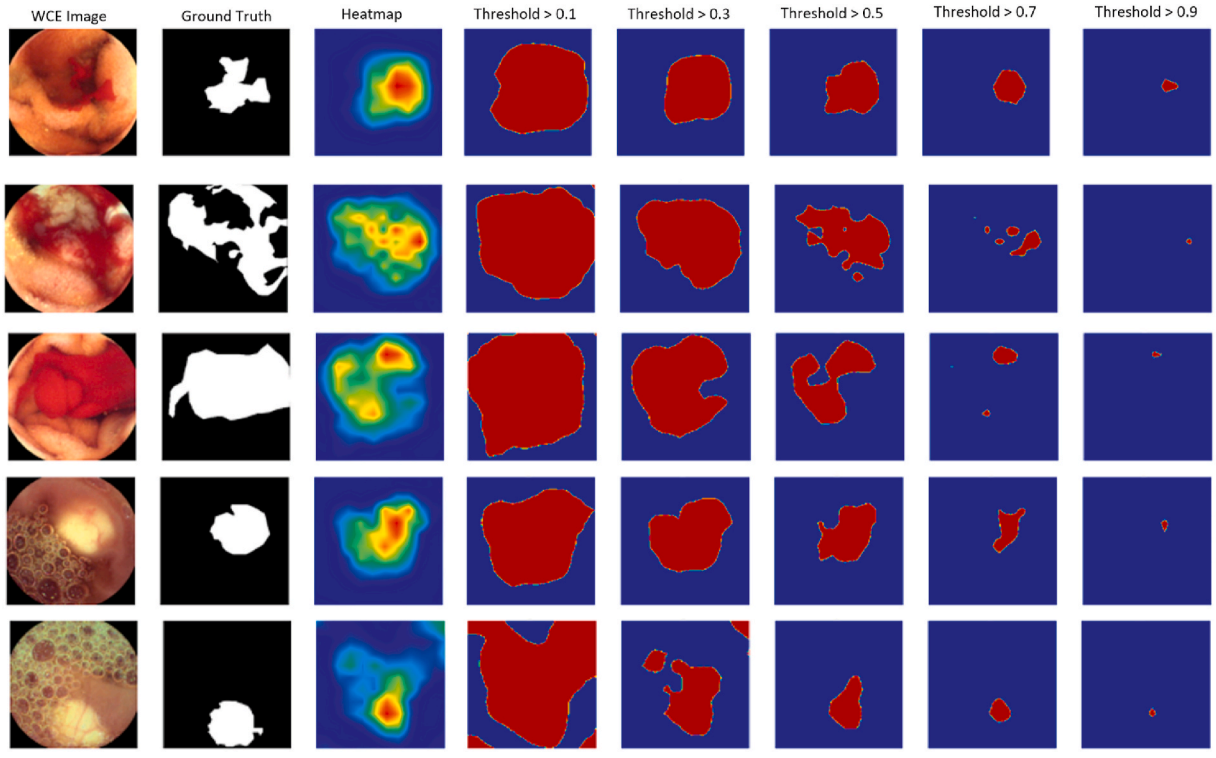

Fig. 11. Some sample segmentation results of Grad-CAM++ using different threshold values.

is derived by combining Eqs. (3) and (7).

$$S^c = \sum_l \left[\sum_i \sum_j \left\{ \sum_a \sum_b \beta_{ab}^{lc} ReLU\left(\frac{\partial S^c}{\partial M_{ab}^l} \right) \right\} M_{ij}^l \right] \tag{8}$$

In Eq. (8), the iterators (i, j) and (a, b) are identical and iterate over entire activation map M^l. Since *ReLU* is a threshold function that allows the gradients to flow-back, we can drop it. Therefore, by taking partial derivative on both the sides of Eq. (8), for a specific class c, an activation map l, and the class score S^c, the gradient weights β_{ij}^{lc} are computed using Eq. (9).

Table 7
Grad-CAM++ segmentation performance with different threshold values.

Threshold	F_wIoU	$MIou$	Dc
0.1	0.73	0.47	0.41
0.2	0.73	0.48	0.43
0.3	0.75	0.53	0.45
0.4	0.75	0.54	0.46
0.5	0.76	0.55	0.47
0.6	0.74	0.54	0.46
0.7	0.74	0.54	0.46
0.8	0.72	0.54	0.45
0.9	0.72	0.54	0.44

Table 8
Localization performance comparison by taking union and intersection of segmentation masks produced by Grad-CAM++ ($M1$) and SegNet ($M2$).

	$M1 \bigcup M2$			$M1 \cap M2$		
Threshold	F_wIoU	$MIou$	D_c	F_wIoU	$MIou$	D_c
0.1	0.70	0.48	0.41	0.83	0.59	0.56
0.2	0.72	0.48	0.44	0.81	0.59	0.55
0.3	0.72	0.49	0.46	0.79	0.57	0.55
0.4	0.73	0.49	0.48	0.78	0.55	0.54
0.5	0.74	0.50	0.50	0.77	0.55	0.54
0.6	0.74	0.50	0.50	0.76	0.55	0.53
0.7	0.74	0.52	0.51	0.76	0.55	0.52
0.8	0.75	0.55	0.51	0.74	0.52	0.50
0.9	0.75	0.55	0.52	0.69	0.47	0.40

$$\beta_{ij}^{lc} = \frac{\frac{\partial^2 S^c}{\left(\partial M_{i,j}^l\right)^2}}{2\cdot\frac{\partial^2 S^c}{\left(\partial M_{i,j}^l\right)^2} + \sum_a\sum_b M_{ab}^l\left\{\frac{\partial^3 S^c}{\left(\partial M_{i,j}^l\right)^3}\right\}} \tag{9}$$

The class-wise saliency maps for a given image I are calculated as the linear combinations of forward activation maps. Each spatial element in the saliency map L^c is then computed using Eq. (10).

$$L_{ij}^c = ReLU\left(\sum_l w_l^c \cdot M_{ij}^l\right) \tag{10}$$

The localization result obtained using Grad-CAM++ is combined with the result of SegNet to get the final segmented image. Pixelwise *AND* operation is performed to compute the final segmented region of an anomaly in a given input image.

4. Experimental results and discussion

In the present section, the performance of the WCENet anomaly detection and localization model is evaluated on the publicly available KID [52] dataset. The model's performance is also compared with some state-of-the-art methods that are relevant to the present work. All the experiments are performed on an Intel Xeon processor with 64 GB RAM and 8 GB Nvidia Quadro P4000 GPU. Keras API 2.1.3 is used with Tensorflow 2.2 as the backend to code the proposed model and all the ML and DL algorithms used in the comparison.

4.1. Dataset

Experiments are performed on the KID dataset [52] which consists of two subsets KID-I [17] and KID-II [11]. The KID-I contains 77 images of various classes of anomalies like angiectasia (27), lymphangiectasia (9), polyps (6), ulcers (9), bleeding (5), stenoses (6), aphthae (5), chylous cysts (8). The KID dataset-II contains images with anomalies of three classes that are polyp (44), vascular (303) and, inflammatory lesion (227), and normal images from different parts of the GI tract like esophagus (282), stomach (599), small bowel (728) and colon (169). The anomaly classes in the KID-I can also be broadly categorized into one of the classes in the KID-II. The abnormal images belonging to angiectasia, lymphangiectasia, bleeding, and stenoses are mapped to the vascular class whereas aphthae and ulcer images are grouped into the inflammatory class. After the categorization of images in the KID-I, they are merged with the KID-II. Images of the size 360 × 360 × 3 are captured using a MicroCam capsule endoscope (IntroMedic Co, Seoul, Korea).

The anomalies belonging to different classes have variations in texture and color as depicted in Fig. 7. The inflammatory class commonly referred to as inflammatory bowel disease (IBD) can be characterized by chronic inflammation in the intestinal walls. It can be a wound-like structure swollen up, sometimes turning red in color. Ulcers are common in IBDs, which appear like a wound with pale yellow or red color. A polyp is another kind of GI disease that is formed on the lining of the colon that looks like a blob of cells. It is formed due to its unregulated growth. In most cases, polyps have a color similar to the intestinal walls but their structure differentiates them from the normal regions. In the GI tract, there can be syndromes with abnormalities in mucosal and submucosal vessels. These vessels may cause bleeding and can be referred to as a vascular anomaly.

The proposed model is trained and tested on the merged KID dataset. The original images in the dataset are of the size 360 × 360 with black borders. In pre-processing steps, the borders are removed and the final images are of size 320 × 320. Sometimes the light source of the capsule gets obstructed by some clinical events [53] resulting in poor quality of frames. To deal with this problem, the contrast limited adaptive histogram equalization (CLAHE) method is used for image contrast enhancement [54]. Since the dataset is imbalanced with the majority of normal images, augmentation is also performed by applying random geometric transformations like rotation between − 20 to + 20°, zooming with a factor of 0.2, horizontal and vertical flips. In addition, random Gaussian noise is used for data augmentation. The number of images in each class before and after augmentations is listed in Table 1. A ratio of 4: 1 is taken for splitting the dataset into training and testing sets.

4.2. Performance of WCENet

The proposed WCENet is trained to classify the WCE images into one of the four classes that are inflammatory, polyp, vascular, and normal as discussed in the previous section. The WCENet classifier is trained using the categorical cross-entropy loss with a learning rate of 0.001, the number of epochs = 200, and Adam optimizer with momentum = 0.9. The performance of the model is tested with and without an attention mechanism and it is found that the attention mechanism provides slightly better results (Table 2).

4.3. Comparative analysis of WCENet

This section provides the experimental results to analyze and compare the performance of WCENet with nine different schemes introduced in Refs. [9,11,18,27,28,38,40,43] and, [55]. Methods involving handcrafted features for the classification of WCE images as well as deep learning techniques are considered for comparative analysis.

1) Deep learning-based methods: WCENet is compared with six different CNN-based classifiers introduced in Refs. [9,11,27,28,43] and [55] (Table 5). These include patch-based CNN models introduced in Refs. [27,28,55] and CNN models that work on full images [9,11,43]. In the patch-based schemes, an image patch is defined to be abnormal if most of its pixels fall in the abnormal region and normal otherwise. For a fair comparison of the performance of WCENet with the patch-based methods, even if the single patch in the image is identified as abnormal then the whole image is

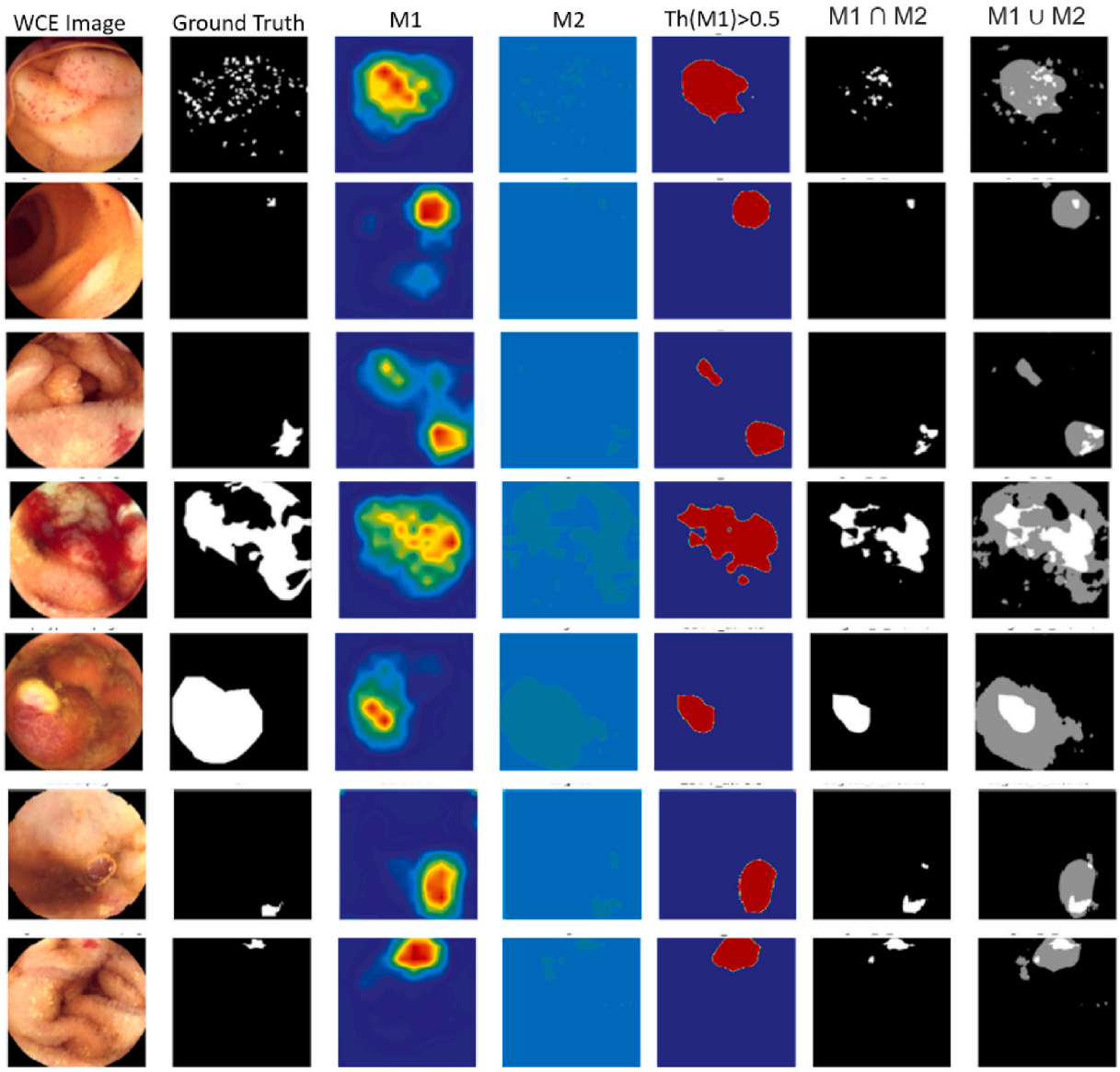

Fig. 12. Localization results produced by taking the union and intersection of segmentation masks produced by GradCAM++(*M*1) and SegNet(*M*2).

considered as abnormal. The CNN model in Ref. [55] is trained on RGB color image patches, whereas $a-$ channel of CIE-Lab color space is used to train the models in Refs. [27,28]. Since the number of abnormal patches is less as compared to the normal patches, a balanced set is prepared for experiments with 5518 patches in each class (c.f [27,28,55]) and the patch size of 32×32 is taken for the experiment. Other than the patch-based models, the CNN architectures in Refs. [9,11,43] work on the full image. The parametric configurations of the above six architectures and the proposed model are given in Table 3.

2)Conventional ML methods: Three conventional ML methods are also compared with WCENet. The first method is by Yuan et al. [40] according to which a bag-of-features is computed making use of SIFT and LBP. Another method is the Color Histogram of Block Statistics (CHOBS) proposed by Ghosh et al. [38] in which some statistical values are calculated over the blocks extracted from the image and on which the color histograms are computed. The third method is by Jain et al. [18] that uses fractal features based on the differential box-counting method which are fed to the random forest classier. The parametric configurations of the above three ML techniques are listed in Table 4.

All the experiments were performed on the merged KID dataset with augmentation mentioned in Section 4.1. Since the number of training and testing samples is quite limited, a 5-fold cross-validation is applied on WCENet as well as on all the methods used in the comparative analysis. Results in Table 5 demonstrate that WCENet yields 98% accuracy. Ghosh et al. [43] have employed transfer learning on AlexNet CNN architecture and the classification accuracy of their model is 96%. In patch-based methods, since an image is divided into patches, it may happen the normal region in a patch dominates the abnormal region, and hence the patch might be categorized as normal. This might be a reason for the lower performance of patch-based CNNs given in Refs. [27,55]. Moreover, the CNN proposed by Jia et al. [22] has fewer convolution layers with a small number of filters that might be the reason for its lower performance with 82% accuracy. We anticipate that the features used in Refs. [18,38,40], have their limitations in identification of a wide range of anomalies in the dataset with different colors, textures, shapes. This could be the reason for the relatively lower performance of these methods. Along with the accuracy and F1-score, the area under the receiver operating characteristic curve abbreviated as AUC is also analyzed. It can be drawn from Table 5 that WCENet performs better than the other state-of-the-art methods with the highest

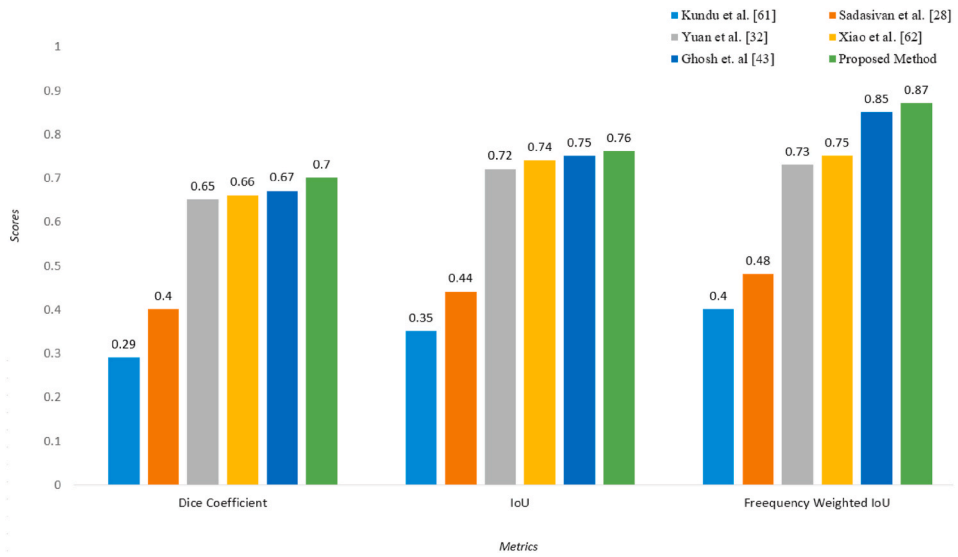

Fig. 13. Segmentation performance of the proposed method and some state-of-the-art bleeding segmentation methods.

Table 9
Performance of WCENet localization on CVC-CLINIC dataset.

WCENet Model	*IoU*	F_wIoU	D_c
Trained with KID dataset	0.48	0.78	0.32
Exclusively trained with CVC-CLINIC	0.90	0.96	0.89

accuracy and AUC score as 98% and 99% respectively. The plots of the ROC curves are shown in Fig. 8.

4.4. Experimental results on anomaly localization

The anomaly localization framework explained in Section 3.2 is a hybrid of two methods: a custom SegNet model with an 8−layer CNN encoder-decoder structure and GradCAM++ using the trained WCENet classifier. The selection of the CNN architecture in SegNet [30] is based on the experimental analysis with two popular segmentation models UNet [56] and PSPNet [57]. All these models use a base CNN architecture. To analyze the performance of segmentation models with respect to the base architecture, three popular CNN models are taken to constitute the base model in UNet, PSPNet, and SegNet. These base models are ResNet50 [58], VGG16 [59], and MobileNetV1 [60]. Apart from this, an 8-layer custom encoder-decoder model is also used for performance evaluation represented in Fig. 5. Three popular metrics are used for comparison namely mean IoU (*MIoU*), frequency weighted intersection over union (F_WIoU), and Dice coefficient (D_c). *MIoU* and F_WIoU can be calculated with the help of Eqs. (11) and (12).

$$MIoU = \frac{1}{N}\cdot\frac{\sum_j n_{jj}}{L_j + \sum_k n_{kj} - n_{jj}}, \tag{11}$$

$$F_wIoU = \frac{1}{\sum_s L_s}\cdot\frac{\sum_j L_j * n_{jj}}{L_j + \sum_k n_{kj} - n_{jj}}, \tag{12}$$

where N is the total number of classes, and n_{kj} is the total number of pixels in the class j identified by the method, but originally belonging to the class k. Further, L_j is the total number of pixels belonging to class j in the ground truth. The dice coefficient is somewhat similar to the F1 score used in the classification. It is computed as the ratio of two times the area of intersection between the ground truth and the predicted segmentation to the union of the ground truth and predicted segmentation areas. Let G

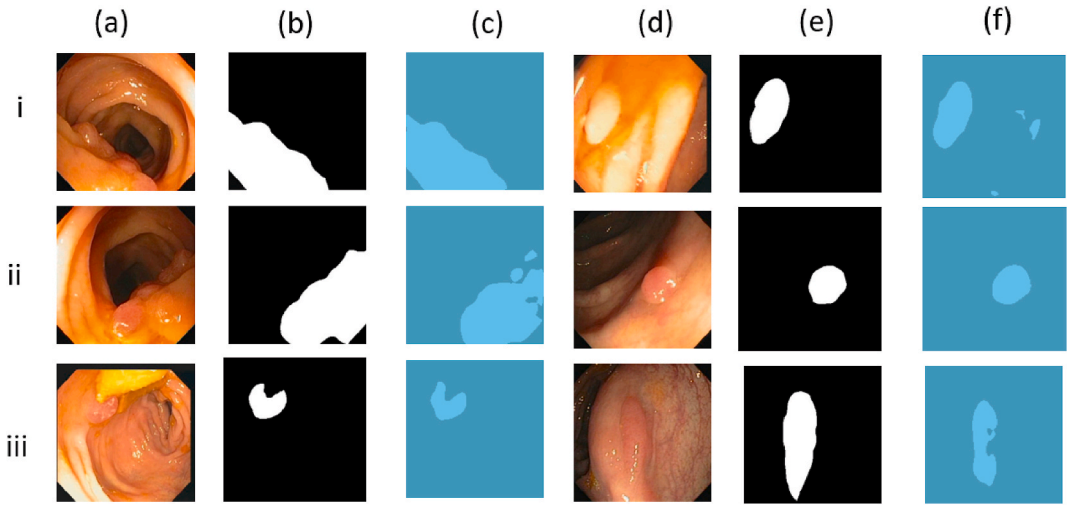

Fig. 14. Segmentation results of SegNet trained on CVC-CLINIC dataset. (a,d). Polyp images in CVC-CLINIC dataset. (b,e) Corresponding ground truths. (c,f) Segmentation results by SegNet.

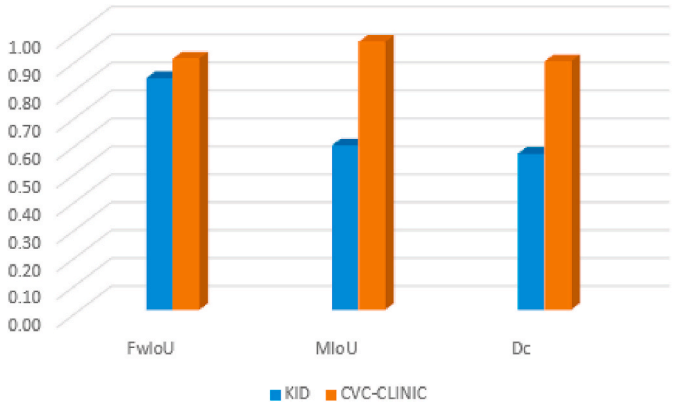

Fig. 15. Quantitative segmentation results of WCENet localization method on both KID and CVC-CLINIC datasets.

be the area of the ground truth and S be the segmented area, then D_c is calculated using Eq. (13).

$$D_c = 2 \times \frac{G \cap S}{|G| + |S|} \tag{13}$$

The segmentation models considered in this study are trained on the same augmented KID dataset, used in the classification stage. All the models are trained for 200 epochs with a categorical cross-entropy loss function. Stochastic gradient descent (SGD) with momentum is used for optimization with the momentum = 0.9 and the learning rate of 0.01.

To find out the best combination of the segmentation model with the appropriate encoder architecture, an empirical study is performed. All the models are trained and evaluated on the augmented KID dataset. The results are recorded in Table 6, which show that the SegNet model with the custom CNN produces better results as compared to other base architectures with the highest F_wIoU of 0.81 as compared to the other encoding architectures. Some of the sample WCE images, their corresponding ground truths, and the segmented regions by the proposed SegNet model are shown in Fig. 9. Although the performance of SegNet with MobileNetV1 is close to SegNet with the custom CNN, we have adopted to use custom CNN due to its lightweight architecture with 8 layers as compared to MobileNetV1 which consists of 13 layers.

The automatic segmentation of WCE images is also performed using GradCAM++ as described in Section 3.2.2. The trained WCENet is utilized for the generation of heatmaps. Some heatmap visualizations are plotted in Fig. 10. The heatmaps are also evaluated on segmentation of anomalies and the qualitative results are shown in Fig. 11. The quantitative results are also reported in Table 7. The colors in the heatmap represent different confidence scores calculated by GradCAM++ through the computation of gradients which can be determined using Eq. (10). The red color signifies the highest confidence of an anomalous region whereas the blue region indicates the normal region. The class C of an image I is identified by referring to the result of the WCENet classifier on the image.

It can be deduced from the individual segmentation results produced by both the techniques discussed above individually that the SegNet

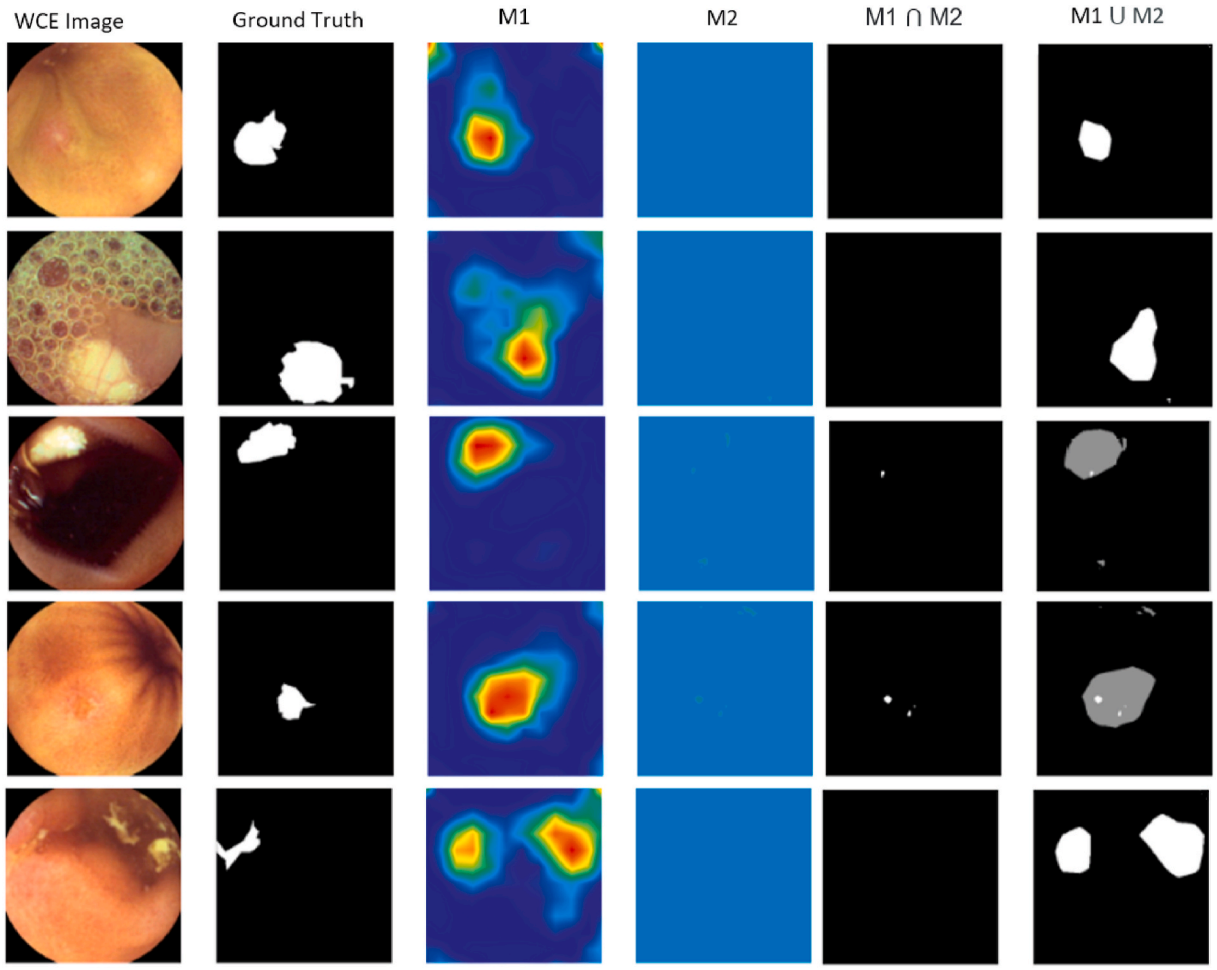

Fig. 16. Failed segmentation results ($M1 \cap M2$) and alternative results by computing union ($M1 \cup M2$) of segmentation masks produced by Grad-CAM++ ($M1$) and SegNet ($M2$).

(M2) technique performs better with a D_c score of 0.55 as compared to Grad-CAM++ (M1) with a D_c score of 0.47. The segmentation results are also computed through the fusion of segmentation masks produced by both the methods. Intersection and union of segmentation masks of both the methods are analyzed as shown in Table 8. It can be seen that the intersection of $M1$ and $M2$ produces a D_c score of 0.56 which is slightly better than individual D_c scores of both M1 and M2. Few examples of the localization results are shown in Fig. 12 for a qualitative comparison.

The localization performance of the proposed model is compared with some state-of-the-art methods. Ghosh et al. [43] have recently reported bleeding segmentation using SegNet [30]. Yuan et al. [32] have suggested a saliency map extraction method in two stages for highlighting bleeding regions, where different color channels are blended in the first stage and a saliency map is obtained in the second stage from the visual contrast by using CIE-Lab and HSV color spaces. Kundu et al. [61] have suggested extracting the bleeding region using inter-plane intensity variation on R–B and R-G planes in the normalized RGB color space. Patch-based CNN is exploited in Ref. [28] where the CNN is trained on normal and abnormal patches. Jia et al. [62] have highlighted bleeding regions by training a fully connected neural network. The results are reported in Fig. 13. Since all these methods are focused on detecting bleeding regions, for a fair comparison of the proposed segmentation method with these methods, only those images are considered that belong to a vascular class of anomaly. It is found that the results obtained by the proposed method are better than other state-of-the-art methods with the highest F_wIoU of 0.87.

To validate the performance of the proposed model trained on the KID dataset [52], it is also tested on another publicly available colonoscopy dataset known as CVC-CLINIC [63]. CVC-CLINIC dataset consists of only polyp frames extracted from the colonoscopy videos. These frames cover a wide variety of polyps. Along with the frames, the dataset also provides ground truths. The dataset contains 612 images taken from 29 different video sequences. Since CVC-CLINIC contains only polyp images, the WCENet classifier could not be trained on the dataset. Therefore, we have tested the performance of the trained WCENet classifier on CVC-CLINIC. It is observed that the WCENet classifier correctly labels 484 frames out of 612 frames as a polyp. The localization capability of the trained WCENet is also tested on the CVC-CLINIC dataset. Further, the custom SegNet model is separately trained on the CVC-CLINIC dataset, and the quantitative results are listed in Table 9. The qualitative results are also shown in Fig. 14. The performance of WCENet on both the datasets augmented KID and CVC-CLINIC is visualized in Fig. 15.

The proposed WCENet performs better in comparison to other methods, but there are situations where the proposed method fails to localize the anomaly. As mentioned earlier, the intersection of the outputs by Grad-CAM++ ($M1$) and SegNet ($M2$) is taken as the final output mask for producing the segmented output. The localization results are quite close to the results produced by SegNet independently. Therefore, SegNet individually can be adopted for localization. But there are cases when SegNet fails to generate the mask. In those cases, we can use the localization mask generated through Grad-CAM++. Also, there can be situations where both the methods generate non-overlapping masks. In such situations, the union of the masks produced by $M1$ and $M2$ may probably help a physician in approximating the probable anomaly regions. In this way, system failure can be minimized. In Fig. 16 some of these cases are demonstrated.

5. Conclusion

In the present paper, a deep CNN model WCENet is proposed for the identification and localization of GI anomalies in WCE images. The model operates in two phases. In the first phase, an input image is passed through an attention-based CNN classifier with 11-layers which classifies the image into one of the four categories namely, inflammatory, polyp, vascular, or normal. If the image is tagged as abnormal, it is passed to the second phase for estimating the anomalous region. Localization network SegNet is supplemented with Grad-CAM++ to produce the localization results. Combining the outcomes of two different localization techniques adds to the reliability of the model. Anomaly localization is the prime requirement for CAD systems in the current scenario of emerging digital healthcare. The proposed model can be applied for the identification and localization of a wide range of anomalies in WCE images. Due to the scarcity of publicly available labeled datasets, experiments are performed only on two datasets, KID [52] and CVC-CLINIC [63]. Comprehensive comparison results with existing approaches demonstrate that the proposed method outperforms other state-of-the-art methods. Although our proposed approach WCENet performs better than legacy methods, there is a room to improve the localization performance in terms of dice score. As future work, more datasets will be considered to yield generalizable results.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgment

This work is partially supported by the project “Prediction of diseases through computer assisted diagnosis system using images captured by minimally-invasive and non-invasive modalities”, Computer Science and Engineering, PDPM Indian Institute of Information Technology, Design and Manufacturing, Jabalpur India (under ID: SPARC-MHRD-231). This work is also partially supported by the project IT4Neuro (degeneration), reg. nr. CZ.02.1.01/0.0/0.0/18 069/0010054 and by the project “Smart Solutions in Ubiquitous Computing Environments”, Grant Agency of Excellence, University of Hradec Kralove, Faculty of Informatics and Management, Czech Republic (under ID: UHK-FIM-GE-2021).

References

[1] J. Dolz, N. Betrouni, M. Quidet, D. Kharroubi, H.A. Leroy, N. Reyns, L. Massoptier, M. Vermandel, Stacking denoising auto-encoders in a deep network to segment the brainstem on mri in brain cancer patients: a clinical study, Comput. Med. Imag. Graph. 52 (2016) 8–18.

[2] M. Habibzadeh, M. Jannesari, Z. Rezaei, H. Baharvand, M. Totonchi, Automatic white blood cell classification using pre-trained deep learning models: resnet and inception, in: Tenth International Conference on Machine Vision (ICMV 2017), vol. 10696, International Society for Optics and Photonics, 2018, 1069612.

[3] T. Rahim, M.A. Usman, S.Y. Shin, A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging, Comput. Med. Imag. Graph. 85 (2020), 101767.

[4] D. Banik, K. Roy, D. Bhattacharjee, M. Nasipuri, O. Krejcar, Polyp-net: a multimodel fusion network for polyp segmentation, IEEE Transactions on Instrumentation and Measurement 70 (2021) 1–12.

[5] B. Li, M.Q.-H. Meng, J.Y. Lau, Computer-aided small bowel tumor detection for capsule endoscopy, Artif. Intell. Med. 52 (1) (2011) 11–16.

[6] A. Karargyris, N. Bourbakis, Wireless capsule endoscopy and endoscopic imaging: a survey on various methodologies presented, IEEE Eng. Med. Biol. Mag. 29 (1) (2010) 72–83.

[7] G. Iddan, G. Meron, A. Glukhovsky, P. Swain, Wireless capsule endoscopy, Nature 405 (6785) (2000), 417–417.

[8] Y. Yuan, J. Wang, B. Li, M.Q.-H. Meng, Saliency based ulcer detection for wireless capsule endoscopy diagnosis, IEEE Trans. Med. Imag. 34 (10) (2015) 2046–2057.

[9] X. Jia, M.Q.-H. Meng, A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images, in: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2016, pp. 639–642.

[10] Y. Yuan, M.Q.-H. Meng, Deep learning for polyp recognition in wireless capsule endoscopy images, Med. Phys. 44 (4) (2017) 1379–1389.

[11] D.K. Iakovidis, S.V. Georgakopoulos, M. Vasilakakis, A. Koulaouzidis, V. P. Plagianakos, Detecting and locating gastrointestinal anomalies using deep learning and iterative cluster unification, IEEE Trans. Med. Imag. 37 (10) (2018) 2196–2210.

[12] A. Novozámskỳ, J. Flusser, I. Tachecí, L. Sulík, J. Bureš, O. Krejcar, Automatic blood detection in capsule endoscopy video, J. Biomed. Opt. 21 (12) (2016), 126007.

[13] G. Lv, G. Yan, Z. Wang, Bleeding detection in wireless capsule endoscopy images based on color invariants and spatial pyramids using support vector machines, in:

2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 6643–6646.
[14] S.A. Karkanis, D.K. Iakovidis, D. Karras, D. Maroulis, Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures, in: Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205) vol. 2, IEEE, 2001, pp. 833–836.
[15] D. Bhattacharjee, A. Seal, S. Ganguly, M. Nasipuri, D.K. Basu, A comparative study of human thermal face recognition based on haar wavelet transform and local binary pattern, Comput. Intell. Neurosci. 2012 (2012), https://doi.org/10.1155/2012/261089. Article ID 261089.
[16] C. Sindhu, V. Valsan, A novel method for automatic detection of inflammatory bowel diseases in wireless capsule endoscopy images, in: 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN), IEEE, 2017, pp. 1–6.
[17] D.K. Iakovidis, A. Koulaouzidis, Automatic lesion detection in wireless capsule endoscopy—a simple solution for a complex problem, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 2236–2240.
[18] S. Jain, A. Seal, A. Ojha, O. Krejcar, J. Bureš, I. Tachecí, A. Yazidi, Detection of abnormality in wireless capsule endoscopy images using fractal features, Comput. Biol. Med. 127 (2020), 104094.
[19] J.-Y. He, X. Wu, Y.-G. Jiang, Q. Peng, R. Jain, Hookworm detection in wireless capsule endoscopy images with deep learning, IEEE Trans. Image Process. 27 (5) (2018) 2379–2392.
[20] X. Xing, Y. Yuan, M.Q.-H. Meng, Zoom in lesions for better diagnosis: attention guided deformation network for wce image classification, IEEE Trans. Med. Imag. 39 (2020) 4047–4059.
[21] S. Jain, A. Seal, A. Ojha, Deep learning models for anomaly detection in wireless capsule endoscopy video frames: the transfer learning approach, in: Smart Computing: Proceedings of the 1st International Conference on Smart Machine Intelligence and Real-Time Computing (SmartCom 2020), 26-27 June 2020, Pauri, Garhwal, Uttarakhand, India,, CRC Press, 2021, p. 423.
[22] X. Jia, M.Q.-H. Meng, Gastrointestinal bleeding detection in wireless capsule endoscopy images using handcrafted and cnn features, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 3154–3157.
[23] M.A. Khan, S. Kadry, M. Alhaisoni, Y. Nam, Y. Zhang, V. Rajinikanth, M.S. Sarfraz, Computer-aided gastrointestinal diseases analysis from wireless capsule endoscopy: a framework of best features selection, IEEE Access 8 (2020) 132850–132859.
[24] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[25] M.I. Razzak, S. Naz, A. Zaib, Deep learning for medical image processing: overview, challenges and the future, in: Classification in BioApps, Springer, 2018, pp. 323–350.
[26] H.-C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers, Deep convolutional neural networks for computer-aided detection: cnn architectures, dataset characteristics and transfer learning, IEEE Trans. Med. Imag. 35 (5) (2016) 1285–1298.
[27] A.K. Sekuboyina, S.T. Devarakonda, C.S. Seelamantula, A convolutional neural network approach for abnormality detection in wireless capsule endoscopy, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 1057–1060.
[28] V.S. Sadasivan, C.S. Seelamantula, High accuracy patch-level classification of wireless capsule endoscopy images using a convolutional neural network, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), IEEE, 2019, pp. 96–99.
[29] Y. Gao, W. Lu, X. Si, Y. Lan, Deep model-based semi-supervised learning way for outlier detection in wireless capsule endoscopy images, IEEE Access 8 (2020) 81621–81632.
[30] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[31] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 839–847.
[32] Y. Yuan, B. Li, M.Q.-H. Meng, Bleeding frame and region detection in the wireless capsule endoscopy video, IEEE.J. Biomed. Health Inf. 20 (2) (2015) 624–630.
[33] T. Rahim, S.A. Hassan, S.Y. Shin, A deep convolutional neural network for the detection of polyps in colonoscopy images, Biomed. Signal Process Contr. 68 (2021), 102654.
[34] B. Li, M.Q.-H. Meng, Automatic polyp detection for wireless capsule endoscopy images, Expert Syst. Appl. 39 (12) (2012) 10952–10958.
[35] D.K. Iakovidis, D.E. Maroulis, S.A. Karkanis, 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.
[36] D.J. Barbosa, J. Ramos, C.S. Lima, Detection of small bowel tumors in capsule endoscopy frames using texture analysis based on the discrete wavelet transform, in: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2008, pp. 3012–3015.
[37] H. Chen, S. Wang, Y. Ding, D. Qian, Saliency-based bleeding localization for wireless capsule endoscopy diagnosis, Int. J. Biomed. Imag. 2017 (2017), https://doi.org/10.1155/2017/8147632. Article ID 8147632.
[38] T. Ghosh, S.A. Fattah, K.A. Wahid, Chobs: color histogram of block statistics for automatic bleeding detection in wireless capsule endoscopy video, IEEE.J. Transl. Eng. Health Med. 6 (2018) 1–12.
[39] S.K. Shah, P.P. Rajauria, J. Lee, M.E. Celebi, Classification of bleeding images in wireless capsule endoscopy using hsi color domain and region segmentation, in: URI-NE ASEE 2007 Conference, 2007.
[40] Y. Yuan, B. Li, M.Q.-H. Meng, Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images, IEEE Trans. Autom. Sci. Eng. 13 (2) (2015) 529–535.
[41] S. Sainju, F.M. Bui, K. Wahid, Bleeding detection in wireless capsule endoscopy based on color features from histogram probability, in: 2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2013, pp. 1–4.
[42] Y. Cong, S. Wang, J. Liu, J. Cao, Y. Yang, J. Luo, Deep sparse feature selection for computer aided endoscopy diagnosis, Pattern Recogn. 48 (3) (2015) 907–917.
[43] T. Ghosh, J. Chakareski, Deep transfer learning for automated intestinal bleeding detection in capsule endoscopy imaging, J. Digit. Imag. (2021) 1–14.
[44] X. Li, H. Zhang, X. Zhang, H. Liu, G. Xie, Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2017, pp. 1994–1997.
[45] E. Ribeiro, A. Uhl, G. Wimmer, M. Häfner, Exploring deep learning and transfer learning for colonic polyp classification, Computational and Mathematical Methods in Medicine 2016 (2016), https://doi.org/10.1155/2016/6584725.
[46] Y. Shin, H.A. Qadir, L. Aabakken, J. Bergsland, I. Balasingham, Automatic colon polyp detection using region based deep cnn and post learning approaches, IEEE Access 6 (2018) 40950–40962.
[47] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning, ArXiv Preprint arXiv:1602.07261.
[48] S. Bianco, R. Cadene, L. Celona, P. Napoletano, Benchmark analysis of representative deep neural network architectures, IEEE Access 6 (2018) 64270–64277.
[49] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[51] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.
[52] A. Koulaouzidis, D.K. Iakovidis, D.E. Yung, E. Rondonotti, U. Kopylov, J.N. Plevris, E. Toth, A. Eliakim, G. Wurm Johansson, W. Marlicz, G. Mavrogenis, A. Nemeth, H. Thorlacius, G.E. Tontini, KID Project: an internet-based digital video atlas of capsule endoscopy for research purposes, Endosc. Int. Open 5 (6) (2017) E477–E483.
[53] S. Seguí, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, J. Vitrià, Generic feature learning for wireless capsule endoscopy analysis, Comput. Biol. Med. 79 (2016) 163–172.
[54] V. Vani, K.M. Prashanth, Color image enhancement techniques in wireless capsule endoscopy, in: 2015 International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), IEEE, 2015, pp. 1–6.
[55] S.V. Georgakopoulos, D.K. Iakovidis, M. Vasilakakis, V.P. Plagianakos, A. Koulaouzidis, Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions, in: 2016 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, 2016, pp. 510–514.
[56] O. Ronneberger, P. Fischer, 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.
[57] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.
[58] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[59] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, ArXiv Preprint arXiv:1409.1556.
[60] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications, ArXiv Preprint arXiv:1704.04861.
[61] A.K. Kundu, S.A. Fattah, M.N. Rizve, An automatic bleeding frame and region detection scheme for wireless capsule endoscopy videos based on interplane intensity variation profile in normalized rgb color space, J. Healthc. Eng. 2018 (2018), https://doi.org/10.1155/2018/9423062. Article ID 9423062.
[62] X. Jia, M.Q.-H. Meng, A study on automated segmentation of blood regions in wireless capsule endoscopy images using fully convolutional networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, 2017, pp. 179–182.
[63] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians, Comput. Med. Imag. Graph. 43 (2015) 99–111.