Medical Image Analysis 68 (2021) 101897

Contents lists available at ScienceDirect

Medical Image Analysis

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

Challenge Report

Toward real-time polyp detection using fully CNNs for 2D Gaussian shapes prediction

Hemin Ali Qadir[a,b,e,*], Younghak Shin[f,**], Johannes Solhusvik[b], Jacob Bergsland[a], Lars Aabakken[d], Ilangko Balasingham[a,c]

[a] *Intervention Centre, Oslo University Hospital, Oslo, Norway*
[b] *Department of Informatics, University of Oslo, Oslo, Norway*
[c] *Department of Electronic Systems, Norwegian University of Science and Technology, Trondheim, Norway*
[d] *Department of Transplantation Medicine, University of Oslo, Oslo, Norway*
[e] *OmniVision Technologies Norway AS, Oslo, Norway*
[f] *Department of Computer Engineering, Mokpo National University, Mokpo, Korea*

ARTICLE INFO

Article history:
Received 11 April 2020
Revised 26 October 2020
Accepted 28 October 2020
Available online 12 November 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Polyp detection
Deep learning
Colonoscopy
Convolutional neural networks
Real-time detection

ABSTRACT

To decrease colon polyp miss-rate during colonoscopy, a real-time detection system with high accuracy is needed. Recently, there have been many efforts to develop models for real-time polyp detection, but work is still required to develop real-time detection algorithms with reliable results. We use single-shot feed-forward fully convolutional neural networks (F-CNN) to develop an accurate real-time polyp detection system. F-CNNs are usually trained on binary masks for object segmentation. We propose the use of 2D Gaussian masks instead of binary masks to enable these models to detect different types of polyps more effectively and efficiently and reduce the number of false positives. The experimental results showed that the proposed 2D Gaussian masks are efficient for detection of flat and small polyps with unclear boundaries between background and polyp parts. The masks make a better training effect to discriminate polyps from the polyp-like false positives. The proposed method achieved state-of-the-art results on two polyp datasets. On the ETIS-LARIB dataset we achieved 86.54% recall, 86.12% precision, and 86.33% F1-score, and on the CVC-ColonDB we achieved 91% recall, 88.35% precision, and F1-score 89.65%.

1. Introduction

Colorectal cancer (CRC) is the third most common cause of cancer mortality for men and women globally, and CRC is the overall second leading cause of cancer-related death (Bray et al., 2018). CRC most often begins as growths of glandular tissue in the mucosal layer of the bowel. Most cases of CRC are initially non-cancerous called polyps. However, if polyps are left untreated, they may become malignant and potentially life-threatening cancer (Arnold et al., 2017). Thus, early detection and removal of pre-cancerous polyps in the colon are crucial for prevention.

Colonoscopy is the most sensitive method for colon screening. It is effective for detection of colonic lesions and polyps of any size, and allows removal of lesions during the procedure. Colonoscopy is an operator-dependent procedure and prone to human errors. Polyp miss rate is reported to be as high as 22%-28% in certain cases (Leufkens et al., 2012). A number of supportive systems have been proposed to help clinicians detect polyps and tumors during colonoscopy, thus reducing polyp miss-rate and optimize the screening procedure.

Deep learning-based detection models which adopt pre-trained deep CNN networks have been successfully applied for automatic polyp detection (Bernal et al., 2017; Shin et al., 2018; Qadir et al., 2019; Qadir et al., 2019; Sornapudi et al., 2019; Wang et al., 2019a; 2019b; Zhang et al., 2019). Most of these models are slow (Yu et al., 2016; Pogorelov et al., 2018; Bernal et al., 2017; Shin et al., 2018; Qadir et al., 2019; Kang and Gwak, 2019) or have difficulty detecting ambiguous types of polyps such as flat-shaped or small polyps (Bernal et al., 2012; 2013; Tajbakhsh et al., 2013; Qadir et al., 2019). A highly accurate supportive system may be crucial to help endoscopists reduce polyp miss rate during colonoscopy. Moreover, a detection system can only be used if it is fast enough for real-time deployment. Most studies have focused on improving

[*] Corresponding author.
[**] Principal corresponding author.
E-mail addresses: hemina.qadir@gmail.com, hqadir2011@my.fit.edu (H.A. Qadir), shinyh0919@gmail.com (Y. Shin).

https://doi.org/10.1016/j.media.2020.101897

detection performance rather than on real-time aspects. In recent years, researchers have become increasingly interested in developing real-time polyp detection systems (Zhang et al., 2018; Mohammed et al., 2018; Wang et al., 2019a; 2019b; Zhang et al., 2019; Liu et al., 2019).

In the colon, there are many polyp-like structures with strong edges, including colon folds, blood vessels, specular lights, luminal regions, air bubbles, etc (Qadir et al., 2019). This is one of the main challenges in the automatic polyp detection task (Shin et al., 2018). When a model is trained to segment polyps from the background, binary masks are used as the ground-truth images, which have very strong outer edges. During training, the binary masks may lead the model to learn edges as one of the strongest features to distinguish polyps. Therefore, such models tend to produce many false positives (FP) (Shin et al., 2018; Qadir et al., 2019).

Most of the CNN-based encoder-decoder models, which are commonly used for object segmentation, can be implemented for real-time applications (Ronneberger et al., 2015) because they are designed to predict a binary mask in a single shot feed-forward fully convolutional neural network (F-CNN), meaning there is no need for a second stage or anchor proposals (Ren et al., 2015; Liu et al., 2016). These models can only predict pixel-wise confidence value and a threshold value is applied to produce the final output binary masks. For object detection, a more explicit mechanism is needed to predict the confidence value for the whole object (Ronneberger et al., 2015). The confidence value is important because a threshold value can be set for the detection confidence to eliminate some FP outputs which tend to have low detection confidence values (Qadir et al., 2019; Shin et al., 2018; Qadir et al., 2019).

In this paper, we aim to use CNN-based encoder-decoder network variants for polyp detection. To tackle the two problems discussed above, we propose to use two-dimensional (2D) Gaussian masks as the ground-truth masks for polyp regions instead of using binary masks, which are normally used to train these types of CNN networks for object segmentation. In this way, we force the CNN networks to predict 2D Gaussian shapes for polyp regions. We propose that 2D Gaussian masks are more efficient than binary masks to reduce the impact of the outer edges during training because a 2D Gaussian shape has smaller values on the tails compared to the values around the mean. This property of the 2D Gaussian shape can give less importance to the outer edges and force the models to learn surface patterns more efficiently than binary masks. The strength of the predicated 2D Gaussian shapes can be used as the confidence values of the detection to further reduce FP outputs.

2. Methods

2.1. Polyp detection as a 2D Gaussian shape

Fig. 1 presents our approach to detect polyps in a one-shot manner. Instead of generating a binary output, we enforce a CNN-based encoder-decoder network to predict a 2D Gaussian shape, $\hat{Y}(x,y) \in [0,1]^{W\times H\times 1}$, for a polyp region in an input RGB image, $I(x,y) \in [R]^{W\times H\times 3}$, where W is the width and H is the height of both $I(x,y)$ and $\hat{Y}(x,y)$.

To train a CNN model for 2D Gaussian shape predictions, we convert the binary ground-truth masks, $f(x,y) \in \{0,1\}^{W\times H\times 1}$, to 2D Gaussian ground-truth masks, $Y(x,y) \in [0,1]^{W\times H\times 1}$, as described in Section 2.2. The 2D Gaussian ground-truth masks can reduce the impact of the outer edges during training, forcing the model to learn not only the outer edges but also other important features of polyps such as surface patterns. They also help to use the strength of the predicted 2D Gaussian shapes as the detection confidence (Zhou et al., 2019).

The output 2D Gaussian shape $\hat{Y}(x,y)$ has the same resolution as the input image $I(x,y)$, i.e., downsampling is not applied on the ground-truth mask $Y(x,y)$ during training the models. In contrast to (Zhou et al., 2019), this elimination of downsampling allows us to ignore:

- computation of the loss for a local offset prediction as there is no need to recover the discretization error.
- the regression for the polyp size as it is calculated from the predict 2D Gaussian shape $\hat{Y}(x,y)$ which has the same size as the input image $I(x,y)$, using size-adaptive standard deviations σ_x and σ_y (Law and Deng, 2018; Zhou et al., 2019) described in Section 2.4.

2.2. Binary masks to 2D Gaussian masks conversion

Usually, for a dataset of polyp images, binary masks $f(x,y) \in \{0,1\}^{W\times H\times 1}$, are provided as the ground-truth images to indicate the location of the polyps. These binary masks are drawn and confirmed by expert clinicians. In the masks, white pixels (1's) correspond to the polyp regions whereas black pixels (0's) correspond to the background. Fig. 2(b) shows a binary mask provided for the polyp shown in Fig. 2(a). We use a 2D elliptical Gaussian kernel expressed in Eq. (1) to convert all the binary masks, $f(x,y)$, in the training dataset to 2D Gaussian masks, $Y(x,y) \in [0,1]^{W\times H\times 1}$,

$$Y = A \cdot \exp\Big(- (a(x - xo)^2 + 2b(x - xo)(y - yo) + c(y - yo)^2), \Big) \tag{1}$$

where A is the amplitude located at the center, (x_o, y_o), of mass in the binary image $f(x,y)$,

$$m_{00} = \sum_x \sum_y f(x,y), \tag{2}$$

$$m_{10} = \sum_x \sum_y x f(x,y), \tag{3}$$

$$m_{01} = \sum_x \sum_y y f(x,y), \tag{4}$$

$$(x_o, y_o) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right). \tag{5}$$

To rotate the output 2D Gaussian masks according to the orientation, θ, of the polyp mask in $f(x,y)$, we set

$$a = \frac{cos^2(\theta)}{2\sigma_x^2} + \frac{sin^2(\theta)}{2\sigma_y^2}, \tag{6}$$

$$b = \frac{-sin(2\theta)}{4\sigma_x^2} + \frac{sin(2\theta)}{4\sigma_y^2}, \tag{7}$$

$$c = \frac{sin^2(\theta)}{2\sigma_x^2} + \frac{cos^2(\theta)}{2\sigma_y^2}, \tag{8}$$

where σ_x and σ_y are the polyp size-adaptive standard deviations (Law and Deng, 2018; Zhou et al., 2019). We compute the orientation, θ, of the mask in $f(x,y)$ as,

$$\theta = \frac{1}{2} \tan^{-1} \left[\frac{2m_{11}}{(m_{20} - m_{02})} \right], \tag{9}$$

$$m_{11} = \sum_x \sum_y (x - x_o)(y - y_o) f(x,y), \tag{10}$$

$$m_{20} = \sum_x \sum_y (x - x_o)^2 f(x,y), \tag{11}$$

$$m_{02} = \sum_x \sum_y (y - y_o)^2 f(x,y). \tag{12}$$

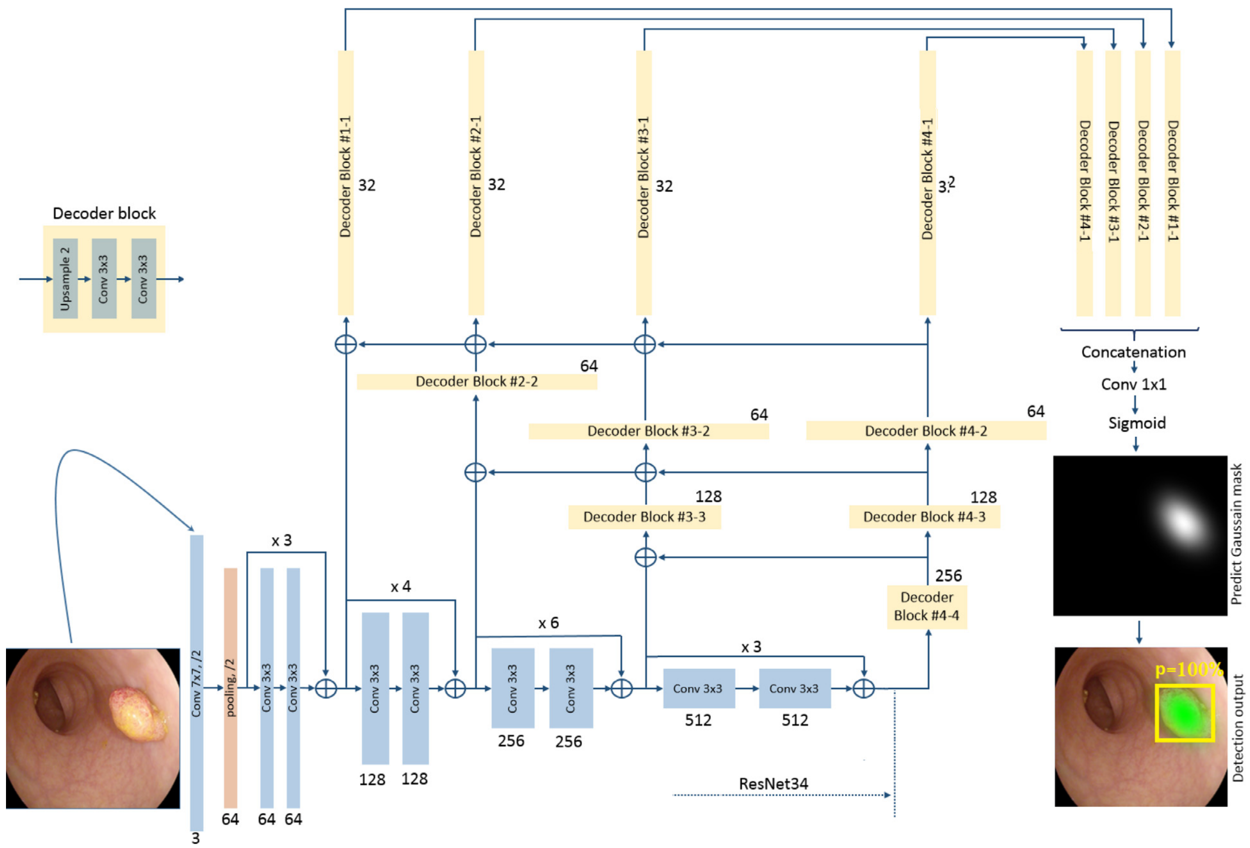

Fig. 1. Our MDeNetplus model for automatic polyp detection. The model is trained on 2D Gaussian masks to predict 2D Gaussian shapes for polyp regions in input images.

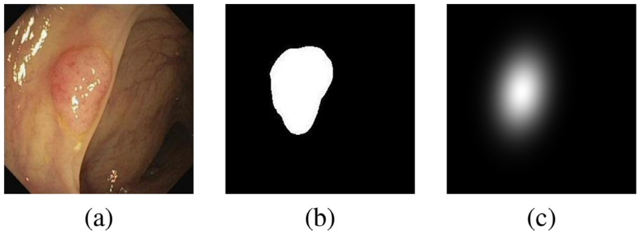
(a) (b) (c)

Fig. 2. An example showing how a binary polyp mask is converted to a 2D Gaussian mask. (a) is the original image with a polyp, (b) the binary mask provided by clinicians, (c) is the 2D Gaussian mask obtained from Eq. (1).

Similar to (Zhou et al., 2019), we set the coefficient $A = 1$, and use it as the confidence value of the detection at the inference time. If two Gaussians overlap, we take the element-wise maximum (Cao et al., 2017). Fig. 2(c) shows a 2D Gaussian mask obtained from Fig. 2(b) using the equations presented above.

2.3. F-CNN models for polyp detection

To prove our concept, we evaluate several different F-CNN based encoder-decoder models, including UNet (Ronneberger et al., 2015), Hourglass (Newell et al., 2016), MDeNet (Qadir et al., 2019), and MDeNetplus—our proposed model. We compare these models for two tasks: 1) polyp segmentation using binary masks as the ground-truth images for training, 2) polyp detection using 2D Gaussian masks as the ground-truth images to force the models to predict 2D Gaussian shapes for polyp regions.

Typically, these models consist of two parts: a contracting path (the encoder) to capture context, and 2) an expanding path (the decoder(s)) that enables precise localization (see Fig. 1). The encoder follows the typical architecture of a CNN with alternating convolution and pooling operations to progressively downsample the resolution and increase the depth of feature maps at every layer. In this study, we use ResNet50 (He et al., 2016) pre-trained on ImageNet database (Deng et al., 2009) as the encoder network for all the models. The decoder(s) gradually up-samples the feature maps at each layer to increase their resolutions and predict an output of the same size as the input RGB image, $I(x, y)$.

UNet (Ronneberger et al., 2015): UNet is developed for medical image segmentation and has proven itself very useful when there is limited amount of data available for training. This network combines up-sampled features maps at the decoder part with the corresponding high-resolution features maps from the encoder part via skip-connections. This feature combination enables precise localization (Ronneberger et al., 2015). For our UNet model, we use AlbuNet34 proposed by (Shvets et al., 2018) for angiodysplasia detection.

EncDec: For the Encoder-Decoder (Enc-Dec) model we use the same architecture of AlbuNet34 without the skip connections.

Hourglass: To build our hourglass model, we stacked two models of AlbuNet34. Hourglass network is famous for yielding the best key-point estimation performance (Newell et al., 2016).

MDeNet: MDeNet is proposed by (Qadir et al., 2019) for semi-automatic polyp annotation. MDeNet consists of an encoder and multiple paths of decoders. Similar to other models, ResNet34 is used as the encoder part to extract different levels of features. At each layer of the encoder, the extracted features are decoded by a decoder. The multiple decoders are meant to increase contextual and semantics information by utilizing the features from different scales and receptive field which helps to segment polyps of different sizes more precisely (Pinheiro et al., 2016; Yu et al., 2018).

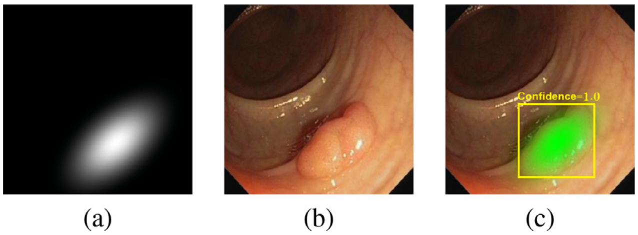

Fig. 3. 2D Gaussian mask (a) is overlaid on the original RGB image (b) and projected back as a bounding box and confidence value shown in (b).

We predict the final output from the outputs of the decoders after concatenating them into a single layer.

MDeNetplus: Our MDeNetplus shown in Fig. 1 is similar to MDeNet with some modifications. Unlike MDeNet, MDeNetplus has feedback connections from decoders of deeper layers to the decoders of previous layers. The feedback connections sum the activation maps of similar layers of different decoders. We prefer summing the activations rather than concatenating them into a single layer to build a smaller network with fewer parameters, helping to realize the network for real-time implantation. This model is based on the concept of aggregation of layers to acquire rich representations that span levels from low to high (Yu et al., 2018), scales from small to large, and resolutions from fine to coarse, iteratively and hierarchically merge the feature hierarchy to make a model with better accuracy.

2.4. From 2D Gaussian shape prediction to bounding boxes and confidence values

At the inference time, we use the peaks in the predicted 2D Gaussian shapes as the confidence values of detection. We calculate the two size-adaptive standard deviations (σ_x and σ_y) for the size of the detection. Fig. 3 shows an example in which the 2D Gaussian shape obtained using Eq. (1) is projected back as a bounding box calculated from σ_x and σ_y and a confidence value (coefficient A) onto the original image. This process allows us to generate all outputs directly from the predicted 2D Gaussian shapes without the need for any post-processing such as IoU-based non-maximum suppression (NMS) (Zhou et al., 2019). This is important to make polyp detection fast for real-time implementation.

3. Experimental details

3.1. Public datasets

To train the models and evaluate their performance, we use three publicly available datasets of polyp images and videos:

1. ETIS-LARIB (Silva et al., 2014): This is a dataset of 196 still images extracted from 34 colonoscopy videos. In total, there are 44 examples of different polyps presented in various sizes and viewpoints. The images have an HD (high definition) resolution of 1225 x 966 pixels. Some images contain two or three polyps, making the total number of polyp appearances 208.
2. CVC-ColonDB (Bernal et al., 2012): This dataset comprises of 300 still images presenting 15 unique polyps coming from 15 different studies. The images have an SD (standard definition) resolution of 574x500. In every image, there exists only one polyp.
3. CVC-ClinicDB (Bernal et al., 2015): This contains 31 unique polyps extracted from 29 colonoscopy videos and presented 646 times in 612 still images with a pixel resolution of 384x288 in SD (standard definition).

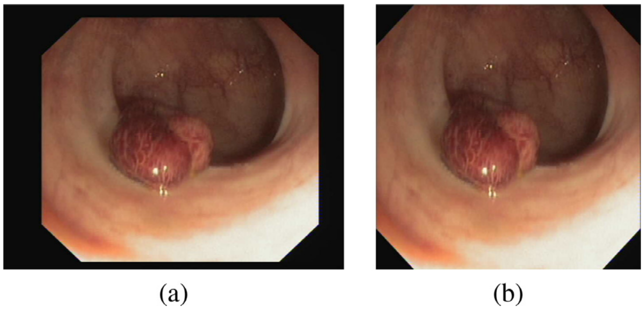

Fig. 4. An example shows that image (a) is cropped to remove the non-informative part as presented in image (b) which is a square image of size 512 x 512 pixels.

In our experiments, we use CVC-ClinicDB for training the models while ETIS-LARIB and CVC-ColonDB are used for the performance evaluation. All three datasets come with ground-truth images in the form of binary masks provided by clinical experts. The ground-truth masks indicate the polyp pixels in the images. The masks are drawn as exact boundaries around the polyp regions.

3.2. Augmentation strategies and preprocessing

We apply several simple pre-processing methods to the input images before used for training the models:

1. Image cropping is applied to remove the canvas around the informative part of the images (see Fig. 4).
2. The input images are resized to 512 × 512 because the pre-trained Resnet34 accepts this image resolution.
3. We re-scale the input images from [0, 255] to [0, 1] and use the mean and standard deviation calculated from the ImageNet dataset to normalize them.

To improve model generalization during training, we apply several image augmentation methods on the fly such as random affine transformations, (e.g., rotation, vertical and horizontal flips), random zoom-in (up to 25%) and zoom-out (up to 50%), and color augmentations in HSV space. Unlike zoom-out, to keep the balance between large and small polyps, we apply zoom-in only up to 25% because the training dataset contains more large polyps than small ones.

3.3. Training the models

We randomly split the training dataset using 5-fold cross-validation to train the models and choose hyper-parameters. We only use images that contain polyps for training. To prevent the models from over-fitting due to shortage of training data, Resnet34 was initialized with ImageNet pre-train weights and the up-sampling layers were randomly initialized. We use Adam optimizer to train the models for 60 epochs with learning rate 0.0001 (chosen using cross-validation) and a batch size of 2 (due to GPU memory restriction).

3.4. Loss functions

It is a known fact that loss function plays an important role in the performance improvement of deep learning. There are many loss functions to choose from and it can be challenging to decide what to pick to obtain the best performance. In this study, we evaluate three loss functions: 1) mean absolute error (L1 loss),

$$L1\ loss = \frac{1}{N}\sum_{i}^{N}|Y_i - \hat{Y}_i|, \tag{13}$$

Table 1
Performance evaluation of the models when trained on Gaussian masks and binary masks.

Model	Gaussian Mask						Binary Mask						MPT (ms)
	TP	FP	FN	Sen %	Pre %	F1%	TP	FP	FN	Sen %	Pre %	F1%	
UNet	174	44	34	83.65	79.81	81.7	165	106	43	79.32	60.88	68.9	31
EncDec	173	45	35	83.17	79.35	81.22	159	116	49	76.44	57.81	65.83	**28**
Hourglass	167	81	41	80.29	67.34	73.25	157	120	51	75.48	56.68	64.74	67
MDeNet	175	34	33	84.13	83.73	83.93	146	97	62	70.19	60.08	64.75	35
MDeNetplus	**177**	**32**	**31**	**85.1**	**84.68**	**84.89**	161	145	47	77.40	52.61	62.64	39

2) mean square error (L2 loss),

$$L2\ loss = \frac{1}{N}\sum_{i}^{N}(Y_i - \hat{Y}_i)^2, \tag{14}$$

3) generative adversarial network (GAN) loss,

$$GAN\ loss = \frac{1}{N}\sum_{i}^{N}\left[logD\Big(concat\,(I_i, Y_i)\Big) + logD\Big(1 - concat\,(I_i, \hat{Y}_i)\Big)\right], \tag{15}$$

where N is the number of samples in the epoch, *concat* is a simple concatenation of I with either Y or $\hat{Y}$, D is the discriminator network, and G is the generator network. For GAN, we use VGG16 (Simonyan and Zisserman, 2014) as the D network to evaluate the output of the G network which can be one of the models discussed in Section 2.3.

3.5. Evaluation metrics

To clinically evaluate a computer-aided diagnosis (CAD), it is important to compute the following medical terminologies:

True Positive (TP): This is a true detection output where the centroid of the detection is located within the polyp masks. Only one is counted if there are multiple overlapped detection outputs for the sample polyp.

True Negative (TN): This is a true detection output where there is no detection for a negative image (image without polyps).

False Positive (FP): This is a false alarm where a wrong detection output is provided for a negative region.

False Negative (FN): This is a false detection output where a polyp is missed in a positive image (image with polyp). We use these terminologies to evaluate the performance of the models in terms of:

Sensitivity (Recall): It measures the ratio of true detection outputs to the total number of polyps in the test dataset. This metric shows the detection ability of a specific model. $Sensitivity\ (Sen) = TP/(TP + FN) \times 100$

Precision: It measures the ratio of true detection outputs to the total number of predicted outputs including false alarms. This metric shows the ability of a model to make correct predictions.

$Precision\ (Pre) = TP/(TP + FP) \times 100$

F-1 score: This metric is clinically important because it shows the balance between sensitivity and precision.

$F1 = (2 * Sen * Pre)/(Sen + Pre) \times 100$

Mean Processing Time per Frame (MPT): It is the actual amount of time needed by a detection model to process a single frame.

4. Results

4.1. Performance comparison of binary and Gaussian masks

We used the ETIS-LARIB dataset and L1 loss to compare Gaussian and binary ground-truth masks on different models. Table 1 shows that Gaussian ground-truth is more efficient and effective than the binary ground-truth. When Gaussian masks were used to train the models to predict 2D Gaussian shapes, all the models were able to detect more TPs and eliminate a number of FPs. These results indicated that our hypothesis of using Gaussian ground-truth is valid. Many FPs could be removed from the final results, because the confidence values (coefficient A) of the predicted masks were less than the threshold value which we set to be 0.5. Many other FPs were eliminated because Gaussian masks were successful for reduction of the effect of outer edges during training.

It can be concluded from Table 1 that MDeNetplus experienced the largest performance improvement with 2D Gaussian masks, especially in terms of precision. The main reason for this superiority is that MDeNetplus hierarchically merges the feature hierarchies to better fuse semantic and spatial information for more accurate detection. This outcome is in line with the results obtained previously (Yu et al., 2018). MDeNetplus was also able to produce fewer FPs because feature aggregation across different layers helps to improve inference of what and where (Yu et al., 2018), making the model well constructed to precisely predict the 2D Gaussian shapes for the polyp regions. However, his method of feature fusion might not be suitable for binary masks because edge information may dominate the features in every decoder of the expanding path, leading to generate more FP outputs. When the network is trained on 2D Gaussian masks, the impact of the edges are reduced and the network more efficiently decodes other types of features to make fewer FP detection outputs and precisely detect more polyps. Fig. 5 presents two examples showing that the MDeNetplus trained on Gaussian masks could precisely predict the location of the polyp without producing FPs, while the same model trained on binary masks produced two FPs along with one correct detection. As can be seen, the two FPs are generated at two locations bounded by some sort of round edges in the image.

We run our tests on an NVIDIA GeForce GTX 1080 Ti to investigate the inference speed of our models. The EncDec model seems to be the fastest model requiring only 28 ms to process a single frame. Compared to other models, the EncDec model has no skip connections and fewer parameters, making it the smallest model. MDeNetplus is the slowest (MTP=39 ms) model with the best performance, but still fast enough for real-time implementation on videos with 25 frames per second.

4.2. Performance evaluation of 2D Gaussian and binary masks on different types of polyp mythologies

In this section, we compare the performance of 2D Gaussian and binary masks in detecting different types of polyps. Based on the morphological shapes, Paris classification divides polyps into several categories: pendunculated (0-Ip), sessile (0-Is), slightly elevated (0-IIa), flat (0-IIb), slightly depressed (0-IIc) and excavated (0-III) (see Fig. 6). ETIS-LARIB dataset contains only pendunculated (0-Ip), sessile (0-Is), and slightly elevated (0-IIa). The sessile and pedunculated polyps are most common types Vleugels et al. (2017). Sessile and slightly elevated polyps lie flat

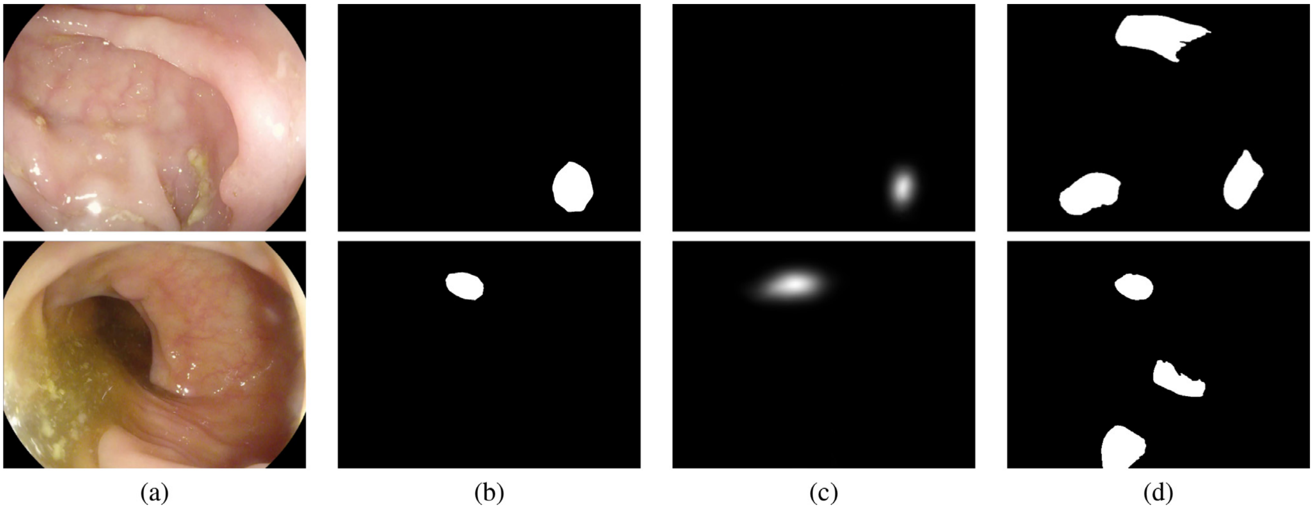

Fig. 5. Two examples presenting the predicted outputs by MDeNetplus model. (a) shows the input images, (b) shows polyp masks drawn by expert clinicians, (c) shows the outputs with no FPs predicted by MDeNetplus when trained on 2D Gaussian masks, (d) shows the outputs contaminated with FPs when MDeNetplus is trained on binary masks.

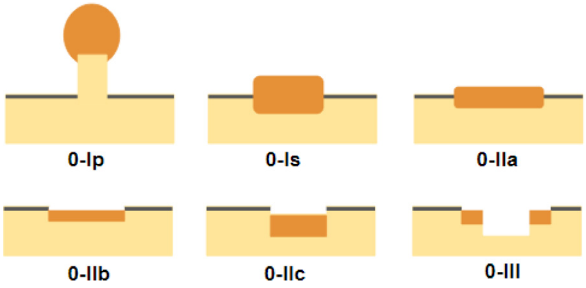

Fig. 6. Paris classification for polyp morphology.

Table 2
Number of missed polyps by 2D Gaussian and binary masks in ETIS dataset.

Types	0-Is	0-Ip	0-IIa
Total no. of polyps	119	29	60
Binary	15	3	29
2D Gaussian	11	3	17

against the surface of the colon's lining, making them harder to detect in CRC screening while pedunculated polyps are mushroom-like tissue growths with a long and thin stalk Vleugels et al. (2017).

In Table 1, we can notice that 16 additional polyps were detected by 2D Gaussian masks than by binary masks. To be exact, we present how many more 0-Is and 0-IIa polyps were detected by 2D Gaussian masks in Table 2. As it can be seen, 2D Gaussian was successful to detect 4 additional sessile and 12 additional slightly elevated polyps. The same 0-Ip polyps were missed by both types of masks. This outcome shows that 2D Gaussian ground-truth was helpful to detect more flat shaped polyps. Fig. 7 presents two 0-IIa polyps (barely noticed by human eyes) detected successfully by our MDeNetplus model trained on 2D Gaussian masks whereas the same model trained on binary masks missed them.

4.3. Comparison of different loss functions

Table 3 shows the performance of MDeNetplus when trained using different loss functions. As seen in the Table, GAN loss is more effective than L1- and L2- loss to force the model to predict 2D Gaussian shapes. We surmise this is because GAN is not only computing the loss between Y and $\hat{Y}$, but also can assess the quality of the predicted Gaussian shapes. If the model predicts an output with irrelevant Gaussian shape, the GAN loss will become large, forcing the model to predict more precise shapes.

Table 3
Performance evaluation of using different loss functions.

loss function	TP	FP	FN	Sen %	Pre %	F1%
L1 loss	177	32	31	85.1	84.68	84.89
L2 loss	174	36	34	83.65	82.85	83.25
GAN loss	**180**	**28**	**28**	**86.54**	**86.12**	**86.33**

4.4. Comparison with other methods on ETIS-LARIB

We followed the same dataset guidelines recommended by endoscopic vision challenge in MICCAI 2015 to train and evaluate our detection models. CVC-ClinicDB is used for training whereas ETIS-LARIB dataset is used for testing. In Table 4, we compare the performance of our best model, MDeNetplus trained with GAN loss, against several state-of-the-art models on ETIS-LARIB dataset. MDeNetplus could outperform the other methods including Faster R-CNN, the-state-of-the-art object detector, in terms of sensitivity (86.54%), and F1 score (86.33%). AFP-Net (Wang et al., 2019a) has 2.42% better precision (88.89%) than our model (86.12%). We surmise this is because they utilized more data to train their model. They used CVC-ClinicVideoDB (Angermann et al., 2017) which comprises 18 videos with a total number of 11,954 frames in which 10,025 frames contain at least a polyp.

Table 4 shows the inference time of the models to process a frame. The fastest model is AFP-Net with only 19 ms of MPT per frame. However, we must mention that they run their model on an NVIDIA GeForce RTX 2080 Ti which is faster than our NVIDIA GeForce GTX 1080 Ti. Nevertheless, we are confident that our MDeNetplus can run faster on an NVIDIA GeForce RTX 2080 Ti.

4.5. Comparison with other methods on CVC-ColonDB

In this experiment, we used CVC-ColonDB to further compare our results with other methods. Table 5 shows that our MDeNetplus trained with GAN was able to produce fewer FP outputs and thus the highest precision (88.35%) and F1 score (89.65%). RCNN-Mask has the highest sensitivity (95.67%) whereas our MDeNetplus

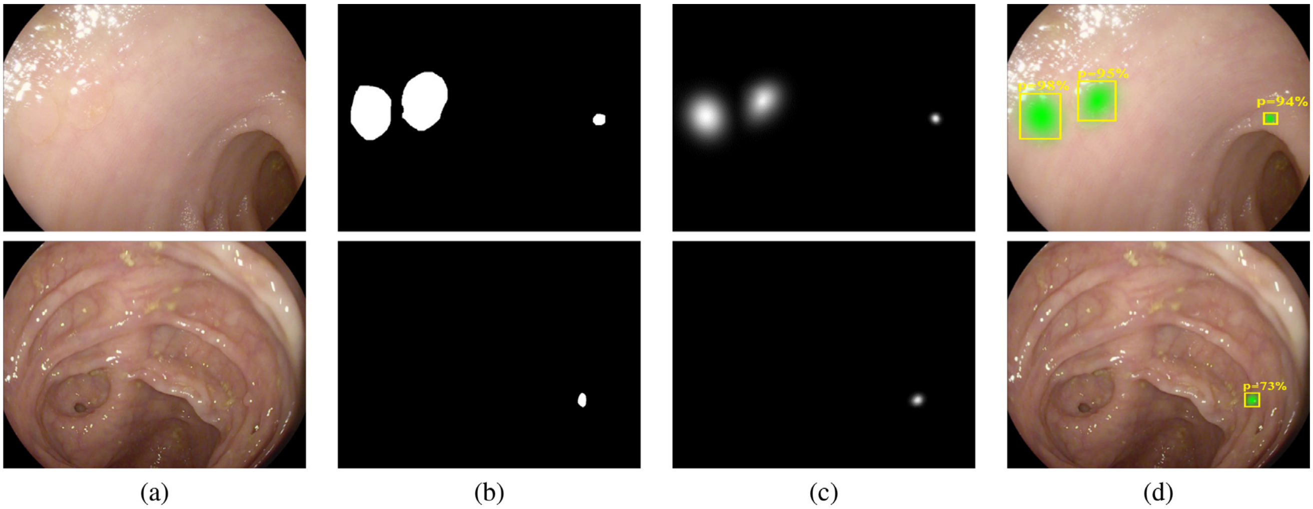

Fig. 7. Two output examples produced by MDeNetplus for slightly elevated polyps in the ETIS-LARIB dataset. The model was able to predict precise 2D Gaussian shapes for all the polyps presented in the two input images. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model.

Table 4
Comparison of Polyp Detection Performance on ETIS-LARIB Dataset.

Methods	Description	TP	FP	FN	Sen %	Pre %	F1%	MPT (ms)
OUS (Bernal et al., 2017)	AlexNet with input patches of 96 × 96	131	57	77	63	69.7	66.1	5000
CUMED (Bernal et al., 2017)	deep contextual network as the backbone	144	55	64	69.2	72.3	70.7	200
Mask R-CNN (Qadir et al., 2019)	Resnet50 as the backbone	N/A	N/A	N/A	72.59	80.0	76.12	430
AFP-Net (Wang et al., 2019a)	anchor free polyp detector	168	21	40	80.77	**88.89**	84.63	**19**
RCNN-Mask (Sornapudi et al., 2019)	R-CNN with Resnet101 +feature pyramid	167	62	41	80.29	72.93	76.43	317
Faster R-CNN (Shin et al., 2018)	Inception-ResNet-v2 as the backbone	167	**26**	41	80.3	81.5	80.9	390
Ensemble Mask R-CNN (Kang and Gwak, 2019)	Two Mask R-CNN models combined	N/A	N/A	N/A	74.37	73.84	N/A	N/A
MDeNetplus	Trained with GAN loss	**180**	28	**28**	**86.54**	86.12	**86.33**	39

Table 5
Comparison of Polyp Detection Performance on CVC-ColonDB Dataset.

Methods	Description	TP	FP	FN	Sen %	Pre %	F1%	MPT (ms)
(Deeba et al., 2020)	WE-SVM	259	256	41	86.33	50.29	56.88	N/A
(Bae and Yoon, 2015)	Discriminative feature learning	212	88	88	70.67	70.67	70.67	637.5
(Bernal et al., 2012)	Valley information	215	241	85	71.67	47.15	56.88	N/A
(Bernal et al., 2013)	Modified valley information	203	90	97	67.77	69.28	68.52	N/A
(Tajbakhsh et al., 2013)	Shape in context	220	90	80	73.33	70.96	72.13	2700
(Sornapudi et al., 2019)	RCNN-Mask with Resnet50	**287**	77	**13**	95.67	78.85	86.58	220
MDeNetplus	Trained with GAN loss	273	**36**	27	91	**88.35**	**89.65**	**39**

has the second-highest (91%) compared to all other methods. However, our MDeNetplus is much faster than RCNN-Mask and needs only 39 ms to process an image. Fig. 8 presents two images in CVC-ColonDB. Again, our method successfully detected a very difficult polyp as shown in the first row of Fig. 8, and even predict the polyp orientation in the image as shown in the second row of Fig. 8. We also encountered FP detection outputs that are shown in Fig. 9. The first row of Fig. 9 shows that MDeNetplus was able to detect the polyp in the input image along with an FP output. The second row of Fig. 9 shows that the model missed the polyp and generated an irregular Gaussian shape in a normal region.

4.6. Effect of resizing the 2D Gaussian and binary masks on the performance

In this experiment, we resized the 2D Gaussian and binary masks to evaluate the effectiveness of smaller and larger masks on the model performance. Fig. 10 shows that when smaller 2D Gaussian masks ($< \sigma$) are used for training the model, sensitivity is low and precision is high because when smaller 2D Gaussian masks are used, less weights are given to the polyp outer edges during training, leading to less FPs being generated for folds and objects with strong edges. When larger 2D Gaussian masks are used, sensitivity increases while precision decreases. From Fig. 10, it can be concluded that the polyp outer edge: a) is an important feature to detect more polyps, b) contributes to produce the majority of FP outputs.

Fig. 11 demonstrates the effect of different sizes of binary masks on model performance. The figure shows that using smaller binary masks (< actual polyp region) are not as effective as using 2D Gaussian shapes to reduce the effect of polyp edges. This is because when smaller binary masks are used, unlike 2D Gaussian masks, part of the polyp region, including the outer edges, are totally excluded from training of the model. It seems that edges cannot be ignored because they are important parts of polyp features. This way of training may fool the model and make it difficult for the model to distinguish between polyp and background. In contrast, 2D Gaussian masks do not totally ignore the edges, but reduce the importance of them by giving them less weights during training of the models.

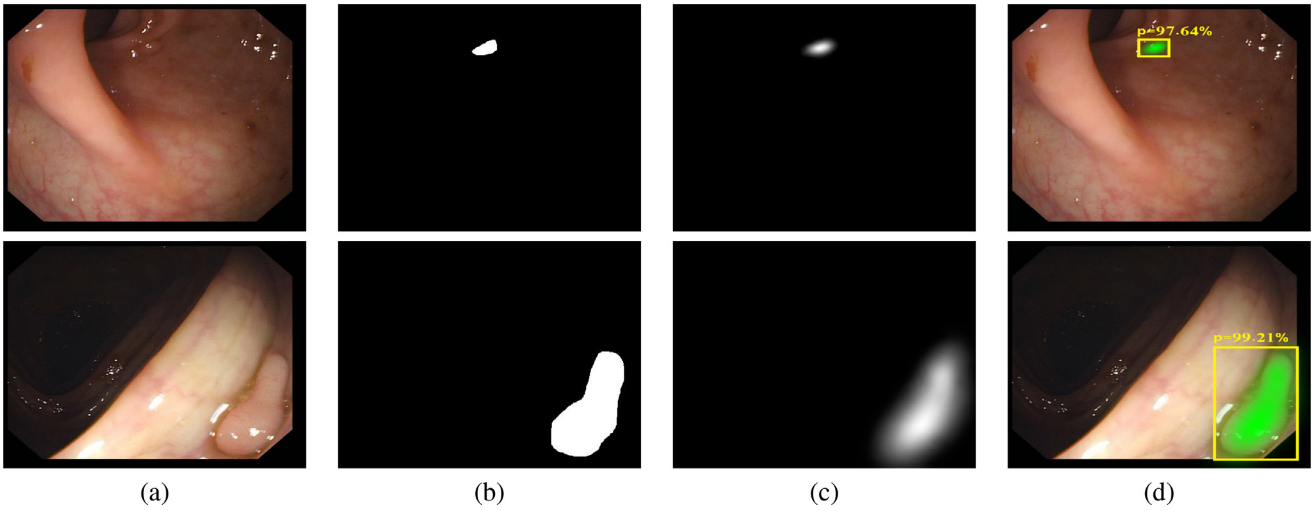

Fig. 8. Two output examples produced by MDeNetplus for input images in CVC-ColonDB. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model.

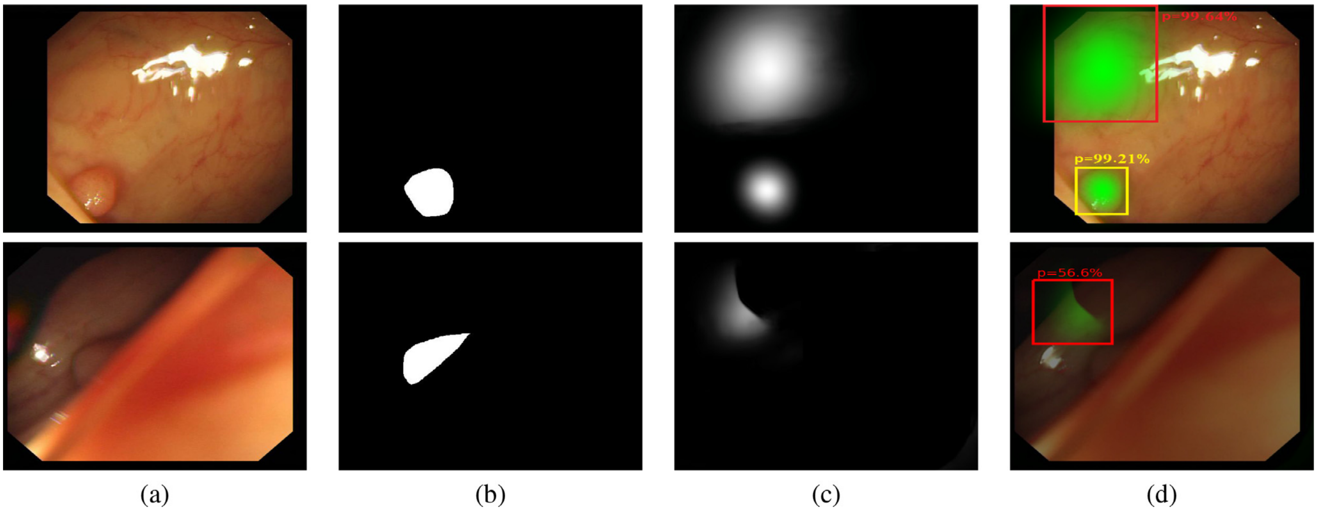

Fig. 9. Examples of FP and FN outputs produced by MDeNetplus for input images in CVC-ColonDB. The yellow bounding box is a TP box while the red bounding boxes are FP outputs. (a) shows the input images, (b) shows the polyp masks drawn by expert clinicians, (c) shows the predicted 2D Gaussian shapes by MDeNetplus model, and (d) is the final detection outputs from the model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

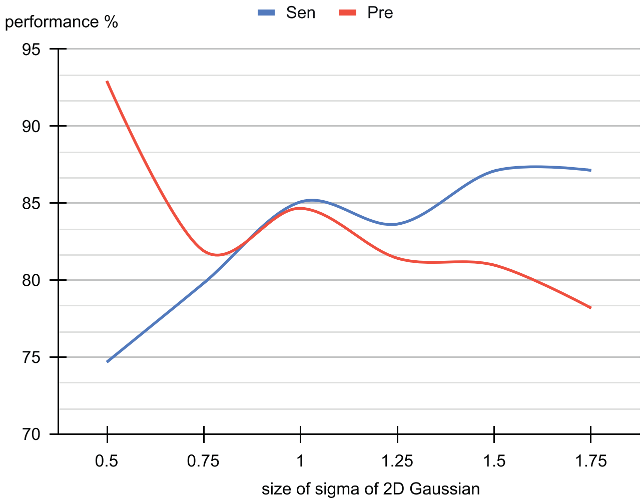

Fig. 10. Effect of resizing 2D Gaussian masks on the model performance.

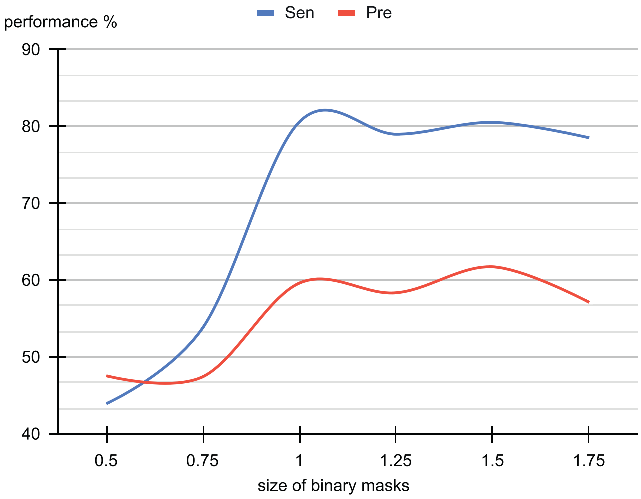

Fig. 11. Effect of resizing binary masks on the model performance.

5. Conclusion

In this paper, we proposed a method for real-time automatic polyp detection with good accuracy. Instead of binary masks, we used 2D Gaussian masks as the ground-truth images to train several convolutional neural networks based encoder-decoder variants which are usually used for object segmentation. We showed that 2D Gaussian masks are more effective and efficient than binary masks to detect more polyps and reduce the number of false positives.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Hemin Ali Qadir: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft. **Younghak Shin:** Validation, Formal analysis, Writing - review & editing. **Johannes Solhusvik:** Project administration, Writing - review & editing. **Jacob Bergsland:** Writing - review & editing. **Lars Aabakken:** Writing - review & editing. **Ilangko Balasingham:** Supervision, Formal analysis, Writing - review & editing.

References

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

Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2017. Global patterns and trends in colorectal cancer incidence and mortality. Gut 66 (4), 683–691.

Bae, S., Yoon, K., 2015. Polyp detection via imbalanced learning and discriminative feature learning. IEEE Trans Med Imaging 34 (11), 2379–2393.

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics 43, 99–111.

Bernal, J., Sánchez, J., Vilarino, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit 45 (9), 3166–3182.

Bernal, J., Sánchez, J., Vilarino, F., 2013. Impact of image preprocessing methods on polyp localization in colonoscopy frames. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, pp. 7350–7354.

Bernal, J., Tajkbaksh, N., Sánchez, F.J., Matuszewski, B.J., Chen, H., Yu, L., Angermann, Q., Romain, O., Rustad, B., Balasingham, I., et al., 2017. Comparative validation of polyp detection methods in video colonoscopy: results from the miccai 2015 endoscopic vision challenge. IEEE Trans Med Imaging 36 (6), 1231–1249.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L., Jemal, A., et al., 2018. Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68 (6), 394–424.

Cao, Z., Simon, T., Wei, S., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299.

Deeba, F., Bui, F.M., Wahid, K.A., 2020. Computer-aided polyp detection based on image enhancement and saliency-based selection. Biomed Signal Process Control 55, 101530.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp. 248–255.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Kang, J., Gwak, J., 2019. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. IEEE Access 7, 26440–26447.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750.

Leufkens, A., Van Oijen, M., Vleggaar, F., Siersema, P., 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 44 (05), 470–475.

Liu, M., Jiang, J., Wang, Z., 2019. Colonic polyp detection in endoscopic videos with single shot detection based deep convolutional neural network. IEEE Access 7, 75058–75066.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: European conference on computer vision. Springer, pp. 21–37.

Mohammed, A., Yildirim, S., Farup, I., Pedersen, M., Hovde, Ø., 2018. Y-net: a deep convolutional neural network for polyp detection arXiv preprint arXiv:1806.01907.

Newell, A., Yang, K., Deng, J., 2016. Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer, pp. 483–499.

Pinheiro, P.O., Lin, T., Collobert, R., Dollár, P., 2016. Learning to refine object segments. In: European Conference on Computer Vision. Springer, pp. 75–91.

Pogorelov, K., Ostroukhova, O., Jeppsson, M., Espeland, H., Griwodz, C., de Lange, T., Johansen, D., Riegler, M., Halvorsen, P., 2018. Deep learning and hand-crafted feature based approaches for polyp detection in medical videos. In: 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS). IEEE, pp. 381–386.

Qadir, H.A., Balasingham, I., Solhusvik, J., Bergsland, J., Aabakken, L., Shin, Y., 2019. Improving automatic polyp detection using cnn by exploiting temporal dependency in colonoscopy video. IEEE Journal of Biomedical and Health Informatics 1-1. doi:10.1109/JBHI.2019.2907434.

Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I., 2019. Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor CNN always perform better? In: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT). IEEE, pp. 1–6.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp. 91–99.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I., 2018. Automatic colon polyp detection using region based deep cnn and post learning approaches. IEEE Access 6, 40950–40962.

Shvets, A.A., Iglovikov, V.I., Rakhlin, A., Kalinin, A.A., 2018. Angiodysplasia detection and localization using deep convolutional neural networks. In: 2018 17th ieee international conference on machine learning and applications (icmla). IEEE, pp. 612–617.

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer. Int J Comput Assist Radiol Surg 9 (2), 283–293.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sornapudi, S., Meng, F., Yi, S., 2019. Region-based automated localization of colonoscopy and wireless capsule endoscopy polyps. Applied Sciences 9 (12), 2404.

Tajbakhsh, N., Gurudu, S.R., Liang, J., 2013. A classification-enhanced vote accumulation scheme for detecting colonic polyps. In: International MICCAI Workshop on Computational and Clinical Challenges in Abdominal Imaging. Springer, pp. 53–62.

Vleugels, J.L., Hazewinkel, Y., Dekker, E., 2017. Morphological classifications of gastrointestinal lesions. Best Practice & Research Clinical Gastroenterology 31 (4), 359–367.

Wang, D., Zhang, N., Sun, X., Zhang, P., Zhang, C., Cao, Y., Liu, B., 2019. Afp-net: realtime anchor-free polyp detection in colonoscopy arXiv preprint arXiv:1909.02477.

Wang, P., Berzin, T.M., Brown, J.R.G., Bharadwaj, S., Becq, A., Xiao, X., Liu, P., Li, L., Song, Y., Zhang, D., et al., 2019. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 68 (10), 1813–1819.

Yu, F., Wang, D., Shelhamer, E., Darrell, T., 2018. Deep layer aggregation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2403–2412.

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2016. Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos. IEEE J Biomed Health Inform 21 (1), 65–75.

Zhang, R., Zheng, Y., Poon, C.C., Shen, D., Lau, J.Y., 2018. Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker. Pattern Recognit 83, 209–219.

Zhang, X., Chen, F., Yu, T., An, J., Huang, Z., Liu, J., Hu, W., Wang, L., Duan, H., Si, J., 2019. Real-time gastric polyp detection using convolutional neural networks. PLoS ONE 14 (3).

Zhou, X., Wang, D., Krähenbühl, P., 2019. Objects as points. arXiv preprint arXiv:1904.07850.