Multi-scale Context-guided Deep Network for Automated Lesion Segmentation with Endoscopy Images of Gastrointestinal Tract

Shuai Wang, Yang Cong, Hancan Zhu, Xianyi Chen, Liangqiong Qu, Huijie Fan, Qiang Zhang, Mingxia Liu

Abstract—Accurate lesion segmentation based on endoscopy images is a fundamental task for the automated diagnosis of gastrointestinal tract (GI Tract) diseases. Previous studies usually use hand-crafted features for representing endoscopy images, while feature definition and lesion segmentation are treated as two standalone tasks. Due to the possible heterogeneity between features and segmentation models, these methods often result in sub-optimal performance. Several fully convolutional networks have been recently developed to jointly perform feature learning and model training for GI Tract disease diagnosis. However, they generally ignore local spatial details of endoscopy images, as down-sampling operations (e.g., pooling and convolutional striding) may result in irreversible loss of image spatial information. To this end, we propose a multi-scale context-guided deep network (MCNet) for end-to-end lesion segmentation of endoscopy images in GI Tract, where both global and local contexts are captured as guidance for model training. Specifically, one global subnetwork is designed to extract the global structure and high-level semantic context of each input image. Then we further design two cascaded local subnetworks based on output feature maps of the global subnetwork, aiming to capture both local appearance information and relatively high-level semantic information in a multi-scale manner. Those feature maps learned by three subnetworks are further fused for the subsequent task of lesion segmentation. We have evaluated the proposed MCNet on 1,310 endoscopy images from the public EndoVis-Ab and CVC-ClinicDB datasets for abnormal segmentation and polyp segmentation, respectively. Experimental results demonstrate that MCNet achieves 74% and 85% mean intersection over union (mIoU) on two datasets, respectively, outperforming several state-of-the-art approaches in automated lesion segmentation with endoscopy images of GI Tract.

Index Terms—Multi-scale Context, Fully Convolutional Network, Lesion Segmentation, Endoscopy Image, Gastrointestinal Tract

I. INTRODUCTION

This study was partly supported by the National Natural Science Foundation of China under Grant (Nos. 61703301 and 61602307), the Natural Science Foundation of Zhejiang Province (No. LY19F020013), the Taishan Scholar Program of Shandong Province in China, the Shandong Provincial Natural Science Foundation (No. ZR2019YQ27), and the Scientific Research Foundation of Taishan University (No. Y-01-2018019).

S. Wang, Y. Cong, L. Qu and H. Fan are with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Nanta Street 114, Shenyang 110016, China. H. Zhu is with School of Mathematics Physics and Information, Shaoxing University, Shaoxing 312000, China. X. Chen is with School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China. Q. Zhang is with School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China. M. Liu is with the Department of Information Science and Technology, Taishan University, Talan 271000, China.

*Corresponding author: M. Liu (mxiuli1226@gmail.com).

AUTOMATIC lesion segmentation of endoscopy images is a fundamental task in computer-aided diagnosis and pathological analysis for gastrointestinal tract (GI Tract) diseases. Currently, various types of endoscopy are designed for GI Tract disease diagnosis, such as colonoscopy [1], [2], gastroscopy [3], [4], narrow-band imaging (NBI) endoscopy [5], zoom-endoscopy [6], and recent wireless capsule endoscopy (WCE) [7], [8]. In addition, advances in medical imaging technology have made it possible to obtain high-resolution endoscopy images, which helps us determine the size and location of lesions in GI Tract. Using endoscopy images, many computer-aided segmentation algorithms have been developed for automated diagnosis of various GI Tract diseases, such as tumor [9], Crohn’s disease [10], bleeding [11], and polyp [12], [13].

Automated lesion segmentation of endoscopy images is a very challenging task due to the poor quality of images, the presence of extraneous matters, blurred boundary between lesion tissue and normal tissues, and diverse appearance in terms of color and texture. As an illustration, we show several representative endoscopy images in Fig. 1. First, due to uneven illumination conditions caused by occlusion between tissues and reflection of body fluids in GI tract, many endoscopy images may have poor quality (e.g., containing dark and bright regions as shown in Fig. 1 (a)), which makes them hard to read. Second, even though each subject is required with an empty stomach before endoscopy, food debris and digestive juice are more or less present in the GI Tract (see Fig. 1 (b)), making it difficult to distinguish internal tissues, folds and organ lumen of GI Tract. Besides, as shown in Fig. 1 (c), the boundary between lesion tissue and nearby normal tissues is usually blurred/unclear due to similar appearance patterns. In addition, as indicated in Fig. 1 (d), endoscopy lesions usually have diverse appearance in terms of color and texture, and these differences in appearance are not only present in different individuals but also in different types of lesions. All these above-mentioned challenges pose great difficulties for automated endoscopy lesion segmentation in GI Tract.

Intuitively, both local and global information of endoscopy images could provide complementary information for lesion segmentation.

Existing methods for automated lesion segmentation in endoscopy images of GI Tract can be roughly divided into two categories: 1) conventional learning-based approaches [14], [15], [16], [17], and 2) deep learning-based approaches [18], [19], [20], [21], [22]. In the first category, hand-crafted fea-
features are often pre-defined for representing endoscopy images, which is independent of the subsequent segmentation model training process. Due to the possible heterogeneity between features and segmentation models, these methods usually result in sub-optimal performance. In the second category, the task of lesion segmentation is usually performed in an end-to-end manner, where feature extraction and segmentation model training are integrated into a unified network. Especially, several fully convolutional networks (FCNs) have been recently developed, which can use an input image with arbitrary sizes and produce an output image with the same size of the input image. Even though FCN methods have shown their advantages over conventional learning-based methods in efficient inference and learning, they usually contain several down-sampling operations (e.g., pooling and convolutional striding) that may result in irreversible information loss of image spatial structures [23], [24], [25], [26]. Considering these challenges in dealing with endoscopy images (e.g., poor quality, presence of extraneous matters, blurred boundary, and diverse appearance), it’s desired to developed advanced methods to capture both the global semantic information and local spatial information of endoscopy images to further boost the performance of FCN-based methods in lesion segmentation.

In this paper, we propose a novel multi-scale context guided deep network (MCNet) to address the lesion segmentation problem in endoscopy images of the gastrointestinal tract (GI Tract), whose schematic illustration is shown in Fig. 2. Different from conventional multi-scale FCN-based methods, our MCNet consists of a global semantic subnetwork to extract the global structure and high-level semantic context and a local detail subnetwork to capture both local appearance and relatively high-level semantic information. The local subnetwork can be easily extended to capture more fine details from different scales. Feature maps generated by the global and local subnetworks are finally fused for lesion segmentation.

The main contributions of this work can be summarized as follows. First, an end-to-end fully convolutional network is developed for endoscopy lesion segmentation in GI Tract. Our network is flexible and can be adapted easily to other applications, which helps reduce manual effort and improve accuracy in real clinical applications. Second, a novel multi-scale context-guided strategy is designed to capture the global semantics as well as the local details in a unified framework. Also, gradients propagate through the global and local paths in an effective and efficient way, which guarantees the convergence of network optimization. Besides, the proposed method has been evaluated on two public datasets (i.e., EndoVis-Ab and CVC-ClinicDB) whose images are collected from volunteers with various lesions in GI Tract. The experimental results demonstrate that our proposed model achieves superior performance with a large margin compared with state-of-the-art methods, especially for images with unclear boundaries between the lesion and normal tissues.

II. RELATED WORK

A. Learning-based Endoscopy Lesion Segmentation

Many conventional learning-based methods have been proposed for automated lesion segmentation of GI Tract endoscopy images [27], [28], [29]. Here, we briefly review the most relevant studies on endoscopy lesion segmentation for automated diagnosis of several typical GI Tract diseases, including bleeding, polyp, ulcer, and tumor.

**Bleeding**: The segmentation of bleeding is very important for clinical diagnosis since bleeding is the main clinical pathology in GI Tract [30]. Li et al. [14] first segmented capsule endoscopy images into superpixels and then classified them as bleeding or not based on the color texture feature. Chu et al. [30] developed a decision support system to predict the bleeding source using eight mathematical models including artificial neural network, support vector machine, k-nearest neighbor, linear discriminant analysis, shrunken centroid, random forest, logistic regression, and boosting. In [31], Fu et al. proposed a computationally complex superpixel-based bleeding segmentation method in the WCE video.

**Polyp**: Automatic polyp segmentation has become a hot research topic for the last 20 years and many outstanding approaches have been proposed [32]. For example, Mamonov et al. [33] assumed that the polyps were characterized as protrusions that were mostly round in shape and a best-fit ball radius was used as a decision parameter of the classifier. Yuan et al. [15] proposed an improved bag of feature method to assist the classification of polyps in WCE images. Bernal et al. [34] presented a validation comparative study on the performance of different polyp segmentation approaches, ranging from methodologies based on hand-crafted methods to trending techniques such as deep learning.

**Ulcer**: As one of the most common symptoms of many serious diseases in GI Tract, many researchers have paid attention to ulcer segmentation. Li et al. [35] made use of curvelet transformation and local binary pattern to design new textural features to distinguish ulcer regions from normal regions. Yu et al. [36] proposed an image processing method based on the bag-of-words model and the feature fusion technique for
ulcer segmentation. Yuan et al. [37] designed a two-stage fully automated computer-aided segmentation system to detect the ulcer. In addition, there are many synergistic frameworks proposed to segment ulcer and other lesions simultaneously. For example, in [38], small bowel polyps, and ulcers in WCE videos were discovered in a synergistic methodology based on the robust and promising log Gabor filters.

**Tumor:** Tumor in GI Tract greatly threatens human health [16], [39]. In order to help clinicians to make an accurate diagnosis, Li et al. [40] proposed color texture features that integrated uniform local binary patterns to characterize WCE images and SVM-based feature selection method to further refine the proposed features for improving the segmentation accuracy. In [17], a set of textural features that integrated multi-scale curvelet and fractal technology were proposed to distinguish normal images from tumor images.

The major disadvantage for these conventional learning-based methods is that they usually rely on hand-crafted features of endoscopy images for lesion segmentation, while feature extraction and segmentation model training are treated as two standalone stages.

**B. Deep Learning-based Lesion Segmentation**

Instead of using the traditional hand-crafted features based on expert knowledge and careful design, deep learning-based methods can automatically learn effective feature hierarchies from the data itself and have shown excellent performance in many fields, such as object recognition [41], image classification [42], [43], [44], [45], and semantic segmentation [46], [18], [19], [47]. In all forms of deep networks, FCN [48], [49], [50] is effective to handle the semantic segmentation problem while it can take arbitrary size input and produce correspondingly sized output by end-to-end learning. FCN has also been widely used for medical image analysis. Christ et al. [20] presented a method to automatically segment liver and lesion in computed tomography (CT) abdomen images using cascaded FCN and dense 3D conditional random fields (CRFs). Ben et al. [21] explored FCN for the task of liver segmentation and liver metastases segmentation in CT examinations. Bi et al. [51] proposed to leverage FCN to automatically segment the skin lesions through a multi-stage segmentation approach in which multiple FCNs learned complementary visual characteristics of different skin lesions. Yuan et al. [22] presented a fully automatic method for skin lesion segmentation by leveraging a 19-layer deep FCN and designing a loss function based on Jaccard distance to eliminate the need of sample re-weighting. Urban et al. [52] tested different CNN architectures for polyp detection. Mahmood et al. [53] predicted depth information and added it to the model training for polyp detection, segmentation, and classification. Wang et al. [54] adopted SegNet [55] as the backbone and integrated with a multi-threaded processing system for polyp segmentation during colonoscopy. Brandao et al. [56] converted three well-known convolutional architectures, e.g., AlexNet [42], GoogLeNet [57], and VGG16 [58], into a fully convolution architecture and fine-tuned their learned representations to identify and segment polyps in colonoscopy images.

However, due to multiple pooling layers and subsampling layers used in FCNs, the prediction is achieved using only
coarse probability maps with large receptive fields obtained by feed-forwarding an input image. Even though the high-level features are able to extract global semantic information, existing FCN-based methods usually result in irreversible information loss of local spatial structures. Furthermore, the use of only large receptive fields in FCNs may result in misdiagnosis of lesions in small regions, especially for endoscopy images that have various appearances in subjects and lesion types. Therefore, it is desired to develop a method to capture both global and local context information in endoscopy images for accurate lesion segmentation in GI Tract. For example, FCN-16s [48] and FCN-8s [48] combined predictions from both the final layer and other pooling layers to let the net predict finer details. Dilation network [59] aggregated multi-scale contextual information by introducing the dilated convolution to expand receptive fields without resolution loss. Zhu et al. [60] designed a coarse-to-fine strategy to refine the results step by step based on the prediction of the previous scale. And there are some existing works learning multi-scale features by inputting multi-scale images [61], [62]. However, these methods usually deal with each scale separately and lack of effective integration of information on different scales.

III. PROPOSED METHOD

As shown in Fig. 2, we propose a novel multi-scale context-guided deep network (MCNet) based on FCN for endoscopy lesion segmentation in GI Tract, which exploits multi-context cues from global semantic features and low-level local detail features. Specifically, our proposed MCNet consists of two types of subnetworks, including one global semantic subnetwork (Gnet) that is designed to exploit the global structure and high-level semantic information, and two local detail subnetwork (Lnet) that is developed to exploit multi-scale appearance information from the shallow layer of Gnet and semantic information from the deep layer of Gnet. These two networks are combined in a cascaded manner, which guarantees gradients to be effectively propagated backward through the whole network. Outputs of these three subnetworks are finally fused together for lesion segmentation, through which multi-scale context information of the input image can be used to guide the network learning.

A. Global Semantic Subnetwork

Fig. 2 shows the basic configuration of our proposed Gnet (gray background), which includes three components, namely, input, reception, and output. The input of Gnet is the Red, Green, and Blue channels of the original image, which is similar to the general FCN framework. The reception component is constructed on the basis of VGG16 network [58]. Specifically, the reception component of our Gnet consists of five convolutional groups, and each convolutional group is followed by a max-pooling layer. These five convolutional groups and max-pooling layers gradually increase the receptive fields and decrease the resolution of generated feature maps. And the final size of feature maps in the reception component is $1/32$ of the original input image.

There are two output branches in Gnet. The first branch contains a convolutional layer and an upsampling layer to obtain the pixel-wise prediction map with the same size of the input image, and each element in the prediction map indicates the probability of a pixel belonging to a specific category (e.g., lesion or normal tissue). The second output branch contains only an upsampling layer to generate feature maps (e.g., with the size of $1/2$ of the input image), and these feature maps are transferred to the following Lnet to compensate for the information loss caused by multiple max-pooling and striding operations in Gnet. That is to say, the proposed Gnet is not only used to extract global context and semantic information for pixel-wise prediction, but also to guide the learning of Lnet through its generated feature maps.

B. Local Detail Subnetwork

After being processed by five convolutional groups and spatial poolings, the output feature maps generated by Gnet have large receptive fields, and thus can effectively extract the global context and semantic information of each input image. On the other hand, using only these low-resolution feature maps in Gnet for pixel-wise dense prediction of lesions will lose the important visual details in each input image, even though we employ an upsampling layer to connect coarse output feature maps and dense pixels in the output probability map. The reason is that those pooling and striding operations in Gnet will lead to irreversible loss of image spatial information.

In order to segment lesion regions in a fine manner, we further design two cascaded local detail subnetworks (Lnet) to capture the local structure information of input images. While Gnet captures the global semantic cues, Lnet mines the local detail cues with adaptive scales in terms of different resolutions. As an illustration, Fig. 2 shows the basic configuration of each Lnet with two scales (blue background). Similar to Gnet, Lnet contains three components: input, reception, and output.

The input component of Lnet contains three types of feature maps: 1) the output feature maps of Gnet or the previous scale (if Lnet has more than one scale and the current scale is not the first one), followed by a $1 \times 1$ convolutional operation; 2) the corresponding feature maps of Gnet with the same resolution to the current scale, followed by a $1 \times 1$ convolutional operation; and 3) the original input image followed by a convolutional operation (with the parameters to ensure that the output has the same resolution as that of the other two input feature maps). In other words, each Lnet perceives the appearance information from semantic information from a deep layer of Gnet or a large-scale Lnet, a shallow layer of Gnet, as well as information conveyed in the original input image. In this way, the proposed Lnet is expected to compensate for those lost context information caused by pooling and striding operations in Gnet. The channel-wise concatenation of these types of input feature maps is used as the input for the following reception component in Lnet. The reception component contains three $3 \times 3$ convolutional layers, without any pooling operation. This can ensure that feature maps generated by these convolutional layers in Lnet have the same resolution, to avoid spatial information loss.
The output component of Lnet-8s (8× upsampling is done to make the output have the same size of the input image) is similar to Gnet, with a two-branch structure to perform pixel-wise prediction and generate feature maps to guide the learning of Lnet-2s at the next scale. Also, the output component of Lnet-2s only contains a convolutional layer and an upsampling layer to obtain the pixel-wise prediction map with the same size of the input image. In this work, we employ only two-scale Lnet (i.e., Lnet-8s and Lnet-2s), while one may extend this network by cascading Lnets with more different scales.

C. End-to-end Endoscopy Lesion Segmentation

Note that Gnet is good at capturing the global structure and high-level semantic features, and Lnet can acquire local details and low-level appearance features. Hence, the proposed MCNet, composed of Gnet and Lnet, can effectively capture information in multi-context. By cascading Gnet, Lnet-8s and Lnet-2s, we can finally obtain the proposed multi-scale context-guided network (MCNet) to capture both global and local context information of each input image. We finally average three probability maps generated by Gnet, Lnet-8s and Lnet-2s, thus obtaining the final segmentation map for the input image.

To optimize the whole network, the multinomial logistic loss is used in the objective function. Specifically, there are four terms: 1) the loss for Gnet \(L_G\), 2) the loss for Lnet-8s \(L_{8s}\), 3) the loss for Lnet-2s \(L_{2s}\), and 4) the loss at the final layer (i.e., used for fusing outputs of different components) of the whole network \(L_F\). Therefore, the proposed objective function is formulated as follows:

\[
L = L_F + \lambda_G L_G + \lambda_{8s} L_{8s} + \lambda_{2s} L_{2s},
\]

where \(\lambda_G\), \(\lambda_{8s}\), and \(\lambda_{2s}\) are parameters to balance the contributions of different terms. In our implementation, we set \(\lambda_G = \lambda_{8s} = \lambda_{2s} = 1\) for an averaged fusion of information in different scales.

D. Implementation

We implement the proposed MCNet based on the Caffe [63] framework. The detailed configuration of MCNet used in this paper is shown in Table I. As shown in Table I, the Lnet has two different architectures corresponding to two scales. All the networks described in this paper are trained and tested in a single Nvidia GTX Titan X GPU with 12G memory. Besides, the learning rate, momentum and weight decay are set to be \(10^{-10}\), 0.95, and 0.0005, respectively.

Recent work in [64] suggests that a pre-trained CNN on a large set of labeled natural images can be transferred from natural to medical images with adequate fine-tuning. Therefore, in this work, we use a pre-trained VGG16 model [58] to initialize network parameters of Gnet and then fine-tune Gnet by backpropagation. Each convolutional layer in Gnet is followed by a Rectified Linear Unit (ReLU).

For Lnet, random Gaussian variables (with the mean value \(u = 0\) and standard deviation \(\delta = 0.001\)) are used for initialization. The standard stochastic gradient descent (SGD) with momentum is employed as the optimization algorithm for network training. To avoid overfitting, each convolutional layer in Lnet-8s and Lnet-2s is followed by a dropout layer and a Parametric Rectified Linear Unit (PReLU) [65].

IV. EXPERIMENT

A. Datasets

In order to verify the performance of our proposed model, we adopt two public datasets to compare with some state-of-the-art FCN frameworks. The first one is EndoVis-Ab, which...
is based on our previous work in collaboration with Chinese PLA General Hospital and has also been released as the data source of MICCAI 2015 Endoscopic Vision Sub-Challenge on Detection of Abnormalities in Gastroscopic Images\(^1\). The original images of EndoVis-Ab are collected from 544 healthy volunteers and 519 volunteers with various lesions, including gastritis, cancer, bleeding, and ulcer. And the original image resolution is \(768 \times 576\). In order to protect personal privacy, we crop each input image to keep the ROI region with the resolution of \(512 \times 416\). Three senior medical experts are invited to annotate the pixel-level ground truth. As a result, EndoVis-Ab contains a total of 389 abnormal images and 309 normal images. Some examples are shown in Fig. 3.

The other one is CVC-ClinicDB \(^2\), which is a dataset of frames extracted from 31 different colonoscopy videos for polyp detection and to be used in the training stage of MICCAI 2015 Endoscopic Vision Sub-Challenge on Automatic Polyp Detection in Colonoscopy Videos. CVC-ClinicDB contains a total of 612 images with resolution as \(384 \times 288\) and each image has its associated manually annotated ground truth covering the polyp. Some examples are shown in Fig. 4.

For EndoVis-Ab, 465 images are used for training and 233 images for testing as indicated in the challenge. For CVC-ClinicDB, 412 images are used for training and 200 images for testing. To achieve stable performance, we perform five-fold cross-validation on the training dataset of EndoVis-Ab and CVC-ClinicDB respectively. For each dataset, all training images are first randomly shuffled and then split into five folds. In each round, four folds are used for training and the remaining one fold for validation.

\(^1\)https://endovissub-abnormal.grand-challenge.org/
\(^2\)https://polyp.grand-challenge.org/CVC-ClinicDB/

1) Three variants of FCN proposed in \(^48\) are implemented, i.e., FCN-32s, FCN-16s, and FCN-8s. These three different architectures differ in the stride of the last convolution (32 pixel stride of FCN-32s, 16 of FCN-16s, and 8 of FCN-8s), which yields different segmentations. Specially, they are all fine-tuned on the pre-trained models in PASCAL VOC 2011.

2) SegNet \(^55\) consists of an encoder network and a corresponding decoder network. Compared with FCN \(^48\), SegNet \(^55\) uses a decoder to upsample the low resolution encoder feature maps to full input resolution feature maps for producing dense predictions. And it is fine-tuned on the pre-trained model in ILSVRC-2014.

3) Dilatation \(^59\) introduces the dilated convolution to expand receptive fields without resolution loss and aggregates multi-scale contextual information, which is fine-tuned on the pre-trained model in PASCAL VOC 2012.

4) Unet \(^67\) has achieved good performance in medical image segmentation, where the skip connection is used to merge the information from different levels. Here, we implement the original Unet \(^67\) and three variations with VGG16, VGG19, and ResNet34 as backbones, respectively \(^68\).

\(^{a}\)\(^{b}\)\(^{c}\)\(^{d}\)\(^{e}\)\(^{f}\)\(^{g}\)\(^{h}\)\(^{i}\)\(^{j}\)\(^{k}\)\(^{l}\)\(^{m}\)\(^{n}\)\(^{o}\)\(^{p}\)\(^{q}\)\(^{r}\)\(^{s}\)\(^{t}\)\(^{u}\)\(^{v}\)\(^{w}\)\(^{x}\)\(^{y}\)\(^{z}\)

C. Experimental Setup

Four metrics used in common semantic segmentation evaluation are reported which are variations on pixel accuracy and region intersection over union (IoU). Let \(n_{ij}\) be the number of pixels of class \(i\) predicted to class \(j\) and let \(t_i = \sum n_{ij}\) be the total number of pixels of class \(i\). When there are \(C\) \((C = 2\) in this work\) different classes, the four metrics are computed as:

- **Pixel-wise accuracy (pACC):** \(\frac{\sum_i n_{ii}}{\sum_i t_i}\)
- **Mean accuracy (mACC):** \(\frac{1}{C} \frac{\sum_i n_{ii}}{t_i}\)
- **Mean IoU (mIoU):** \(\frac{1}{C} \sum_i \frac{n_{ii}}{(t_i + \sum_j n_{ji} - n_{ii})}\)
- **Frequency weighted IoU (fIoU):** \(\frac{1}{\sum_i t_i} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}\)

Note that pACC reports the percent of pixels which were correctly classified, mACC is simply the average over the accuracy of all classes, mIoU is simply the average over the IoU of all classes, and fIoU is an extension of mIoU in which weights are assigned according to the frequency of each class.

D. Results on EndoVis-Ab

Table II gives the performances of our proposed MCNet and compared methods on the test set of EndoVis-Ab. It can be seen that the proposed MCNet achieves the best performance in three metrics except for mACC.

In Fig. 5, we show some qualitative endoscopy lesion segmentation results in the test set of EndoVis-Ab obtained by different methods. The first image is normal with no visible lesion, and the other five images are with different lesions, such as ulcer, polyp, and cancer. We can see that the appearance and size of lesion regions are diverse, and there is a lack of clear structural boundaries between normal regions and lesion regions. Moreover, some regions with poor quality make the data hard to analyze. These all bring obstacles to segment abnormal regions from normal regions even without considering different types of diseases.
For the normal image (the first row in Fig. 5), FCN-8s [48], SegNet [55], Unet [67], and Unet (VGG16) [68] segment some abnormal regions and result in a misdiagnosis. The images from second to sixth rows are with abnormal information that is difficult to distinguish, all compared methods have the situation that missing segmentation for abnormal regions. For FCN-32s [48] and FCN-16s [48], their class score maps are predicted by feature maps with the larger receptive field by an upsampling layer. Therefore, they can effectively capture the large abnormal regions but maybe miss some small abnormal regions. For FCN-8s [48], SegNet [55], Dilation [59], Unet [67], Unet (VGG16) [68], and Unet (VGG19) [68] due to uneven appearance within abnormal regions, the predictions bring many holes and spots, which make abnormal regions incomplete. As a contrast, the proposed MCNet works well on these situations by integrating global semantic structures and local image details.

Table III reports the quantitative results of different methods on CVC-ClinicDB. We can see that the proposed MCNet achieves the best performance in all metrics by a large margin.

<table>
<thead>
<tr>
<th>Method</th>
<th>pACC</th>
<th>mACC</th>
<th>mIoU</th>
<th>fIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-32s [48]</td>
<td>0.929</td>
<td>0.837</td>
<td>0.707</td>
<td>0.813</td>
</tr>
<tr>
<td>FCN-16s [48]</td>
<td>0.896</td>
<td>0.820</td>
<td>0.712</td>
<td>0.822</td>
</tr>
<tr>
<td>FCN-8s [48]</td>
<td>0.869</td>
<td>0.689</td>
<td>0.608</td>
<td>0.771</td>
</tr>
<tr>
<td>SegNet [55]</td>
<td>0.863</td>
<td>0.856</td>
<td>0.709</td>
<td>0.773</td>
</tr>
<tr>
<td>Dilation [59]</td>
<td>0.872</td>
<td>0.853</td>
<td>0.708</td>
<td>0.808</td>
</tr>
<tr>
<td>Unet [67]</td>
<td>0.890</td>
<td>0.862</td>
<td>0.692</td>
<td>0.830</td>
</tr>
<tr>
<td>Unet (VGG16) [68]</td>
<td>0.879</td>
<td>0.823</td>
<td>0.699</td>
<td>0.809</td>
</tr>
<tr>
<td>Unet (VGG19) [68]</td>
<td>0.888</td>
<td>0.859</td>
<td>0.729</td>
<td>0.827</td>
</tr>
<tr>
<td>Unet (ResNet34) [68]</td>
<td>0.869</td>
<td>0.847</td>
<td>0.736</td>
<td>0.804</td>
</tr>
<tr>
<td>MCNet (Ours)</td>
<td>0.909</td>
<td>0.840</td>
<td>0.740</td>
<td>0.841</td>
</tr>
</tbody>
</table>

V. DISCUSSION

In this section, we first perform ablation study to investigate the effectiveness of using multi-scale context information, as well as the influence of the proposed fusion strategy and different trade-off parameters in the loss function. We then analyze the computational complexity of the proposed method and those competing methods, and present the limitations of the current study as well as several future research directions.

A. Effectiveness of the Multi-Scale Cascaded Structure

The proposed MCNet model consists of two subnetworks (i.e., Gnet and Lnet), and Lnet is made up of two scales (i.e., Lnet-8s and Lnet-2s). Note that Gnet and Lnet are not independent, since they communicate with each other and the final segmentation results are obtained by fusing the results of all subnetworks and scales. In order to analyze the effectiveness and necessity of each subnetwork and each scale, we employ the above-mentioned four metrics to evaluate the output pixel-wise prediction score maps generated by our MCNet and its two variants, i.e., 1) MCNet using only a single Gnet (denoted as “Gnet”) with only the second term in Eq. 1; and 2) MCNet using both Gnet and Lnet-8s (denoted as “GLnet-8s”) with the first 3 terms in Eq. 1. The quantitative results on the EndoVis-Ab and CVC-ClinicDB datasets are shown in Fig. 7. Representative segmentation results in the test set are shown in Fig. 8 (a) and Fig. 8 (b), respectively.

From Fig. 7, we can see that when more discriminative information (global or local) is fused into the network, the quantitative results related to all metrics are getting better and better. From Fig. 8, we can see that our MCNet combined with both global and local detail features, the network can gradually refine the results to reduce the misdiagnosis.

Moreover, we show several individual segmentation maps of Gnet, Lnet-2s, and Lnet-8s to demonstrate the role of each scale in Fig. 9. We can see that there exist significant differences between the prediction of each scale and the final results are improved by fusing the learned information from different scales to correct some false predictions of each individual scale.

B. Effectiveness of Fusion Strategy

In our implementation, the three probability maps from different scales are averagely fused as the final output for optimization, namely $L_F$ in Eq. (1). To verify its effectiveness,
Fig. 5. Lesion segmentation results of different methods for six examples from the EndoVis-Ab test set. Here, lesion regions are marked with red color and the normal pixels and backgrounds are marked as black.

Fig. 6. Lesion segmentation results of different methods in the CVC-ClinicDB test set. Here, lesion regions are marked with red color and the normal pixels and backgrounds are marked as black.

Fig. 7. Quantitative results using different components on the (a) EndoVis-Ab and (b) CVC-ClinicDB datasets.
Fig. 8. Pixel-wise prediction results achieved by our MCNet using different components for (a) five subjects from EndoVis-Ab and (b) five subjects from CVC-ClinicDB.

Fig. 9. Pixel-wise prediction results achieved by different scales of our MCNet on EndoVis-Ab.

Fig. 10. Exemplars of segmentation results with unsatisfactory boundary prediction.

**TABLE IV**

<table>
<thead>
<tr>
<th>Method</th>
<th>pACC</th>
<th>mACC</th>
<th>mIoU</th>
<th>fIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>without $L_F$</td>
<td>0.893</td>
<td>0.834</td>
<td>0.729</td>
<td>0.837</td>
</tr>
<tr>
<td>with $L_F$</td>
<td>0.909</td>
<td>0.840</td>
<td>0.740</td>
<td>0.841</td>
</tr>
</tbody>
</table>

we compare the performance of our network supervised under the loss function without and with $L_F$, respectively. For the case without $L_F$, the final segmentation map is obtained by averaging the three predictions from different scales.

From Table IV, we can see that the overall performance will be slightly reduced if without $L_F$, which is mainly because $L_F$ can better fuse information from different scales and then guide the whole network optimization.

**C. Sensitivity Analysis of Loss Weight**

As indicated in Eq. (1), we optimize the whole network using a loss function with multiple terms and three predefined parameters ($\lambda_G$, $\lambda_8s$, and $\lambda_2s$) are used to balance different terms. To verify the sensitivity of our model to these parameters, we have trained two different models separately to highlight the role of global and local terms, one with $\lambda_G = 1$ and $\lambda_8s = \lambda_2s = 0.1$ and the other one with $\lambda_G = 0.1$ and $\lambda_8s = \lambda_2s = 1$. The performance of them and also the original implementation with $\lambda_G = \lambda_8s = \lambda_2s = 1$ is reported in Table V.

From Table V, there is not much difference between the overall performance under different configurations which demonstrates the robustness of our method and that the adopted fusion module $L_F$ can effectively combine the information from different scales.
### Table V
Quantitative results achieved by our method with different loss weight configurations on EndoVis-Ab.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>pACC</th>
<th>mACC</th>
<th>mIoU</th>
<th>fIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>λC = 1, λS = λ2s = 0.1</td>
<td>0.904</td>
<td>0.851</td>
<td>0.736</td>
<td>0.830</td>
</tr>
<tr>
<td>λC = 0.1, λS = λ2s = 1</td>
<td>0.901</td>
<td>0.836</td>
<td>0.754</td>
<td>0.837</td>
</tr>
<tr>
<td>λC = λS = λ2s = 1</td>
<td>0.909</td>
<td>0.840</td>
<td>0.740</td>
<td>0.841</td>
</tr>
</tbody>
</table>

#### D. Computational Complexity

Here, we report the number of parameters, averaged inference time, and used GPU information of each compared method for analysis in Table VI. We can make the following two conclusions: 1) more parameters usually lead to more inference time; 2) the averaged inference time of our method is only slightly different from other methods with a similar number of parameters while our method can achieve more excellent performance.

<table>
<thead>
<tr>
<th>Method</th>
<th>No. of parameters</th>
<th>Inference time</th>
<th>Used GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCN-32s [48]</td>
<td>134,272,704</td>
<td>0.112s</td>
<td>Titan X with 12GB memory</td>
</tr>
<tr>
<td>FCN-16s [48]</td>
<td>134,261,504</td>
<td>0.118s</td>
<td></td>
</tr>
<tr>
<td>FCN-8s [48]</td>
<td>134,259,008</td>
<td>0.112s</td>
<td></td>
</tr>
<tr>
<td>SegNet [55]</td>
<td>10,422,181</td>
<td>0.043s</td>
<td></td>
</tr>
<tr>
<td>Dilation [59]</td>
<td>2,175,597</td>
<td>0.019s</td>
<td></td>
</tr>
<tr>
<td>Unet [67]</td>
<td>2,399,973</td>
<td>0.024s</td>
<td></td>
</tr>
<tr>
<td>Unet (VGG16) [68]</td>
<td>23,752,273</td>
<td>0.037s</td>
<td></td>
</tr>
<tr>
<td>Unet (VGG19) [68]</td>
<td>29,061,969</td>
<td>0.039s</td>
<td></td>
</tr>
<tr>
<td>Unet (ResNet34) [68]</td>
<td>24,456,154</td>
<td>0.035s</td>
<td></td>
</tr>
<tr>
<td>MCNet (Ours)</td>
<td>26,083,264</td>
<td>0.041s</td>
<td></td>
</tr>
</tbody>
</table>

### Table VI
Number of parameters, inference time, and used GPU of each compared method.

**REFERENCES**


