

ELNet:Automatic classification and segmentation for esophageal lesions using convolutional neural network.

Zhan Wu, Rongjun Ge, Minli Wen, Gaoshuang Liu, Yang Chen, Pinzheng Zhang, Xiaopu He, Jie Hua, Limin Luo, Shuo Li

▶ **To cite this version:**

Zhan Wu, Rongjun Ge, Minli Wen, Gaoshuang Liu, Yang Chen, et al.. ELNet:Automatic classification and segmentation for esophageal lesions using convolutional neural network.. Medical Image Analysis, 2021, 67, pp.101838. 10.1016/j.media.2020.101838 . hal-02998748

HAL Id: hal-02998748
https://hal.science/hal-02998748

Submitted on 19 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal Pre-proof

Highlights:

Automatic and accurate esophageal lesion classification and segmentation is of great significance to clinically estimate the lesion status of esophageal disease and make suitable diagnostic schemes. Due to individual variations and visual similarities of lesions in shapes, colors and textures, current clinical methods remain subject to potential high-risk and time-consumption issues. In this paper, we propose an Esophageal Lesion Network (ELNet) for automatic esophageal lesion classification and segmentation using deep convolutional neural networks (DCNNs). The underlying method automatically integrates dual-view contextual lesion information to extract global features and local features for esophageal lesion classification of four esophageal image types (Normal, Inflammation, Barrett, and Cancer) and proposes lesion-specific segmentation network for automatic esophageal lesion annotation of three esophageal lesion types at pixel level. For established clinical large-scale database of 1051 white-light endoscopic images, ten-fold cross-validation is used in method validation. Experiment results show that the proposed framework achieves classification with sensitivity of 0.9034, specificity of 0.9718 and accuracy of 0.9628, and the segmentation with sensitivity of 0.8018, specificity of 0.9655 and accuracy of 0.9462. All of these indicate that our method enables an efficient, accurate and reliable esophageal lesion diagnosis in clinical.

The main contributions of our work can be generalized as follows:

1 For the first time, proposed ELNet enables an automatically and reliably comprehensive esophageal lesions classification of four esophageal lesion types (Normal, Inflammation, Barrett, and Cancer) and lesion-specific segmentation from clinically white-light esophageal images to make suitable and repaid diagnostic schemes for clinicians.

2 A novel Dual-Stream network (DSN) is proposed for esophageal lesion classification. DSN automatically integrates dual-view contextual lesion information using two CNN streams to complementarily extract the global features from the holistic esophageal images and the local features from the lesion patches.

3 Lesion-specific esophageal lesion annotation with Segmentation Network with Classification (SNC) strategy is proposed to automatically annotate three lesion types (Inflammation, Barrett, Cancer) at pixel level to reduce the intra-class differences of esophageal lesions.

4 A clinically large-scale database esophageal database is established for esophageal lesions classification and segmentation. This database includes 1051 white-light esophageal images, which consists of endoscopic images in four different lesion types. Each image in this database has a classification label and its corresponding segmentation annotation.

Journal Pre-proof

Journal Pre-proof

ELNet：Automatic Classification and Segmentation for Esophageal Lesions using Convolutional Neural Network

Zhan Wu[1,†], Rongjun Ge[2,†], Minli Wen[1], Gaoshuang Liu[3], Yang Chen[1,2,4,5,*], Pinzheng Zhang[2], Xiaopu He[3,*], Jie Hua[3], Limin Luo[1,2,4,5], and Shuo Li[6]

1. School of Cyberspace Security, Southeast University, Nanjing, Jiangsu, China
2. School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu, China
3. Department of Gastroenterology, the First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu, China
4. Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing, China
5. Centre de Recherche en Information Biomedicale Sino-Francais (LIA CRIBs), Rennes, France
6. Department of Medical Imaging, Western University, London, Canada

Abstract—Automatic and accurate esophageal lesion classification and segmentation is of great significance to clinically estimate the lesion statuses of the esophageal diseases and make suitable diagnostic schemes. Due to individual variations and visual similarities of lesions in shapes, colors, and textures, current clinical methods remain subject to potential high-risk and time-consumption issues. In this paper, we propose an Esophageal Lesion Network (ELNet) for automatic esophageal lesion classification and segmentation using deep convolutional neural networks (DCNNs). The underlying method automatically integrates dual-view contextual lesion information to extract global features and local features for esophageal lesion classification and lesion-specific segmentation network is proposed for automatic esophageal lesion annotation at pixel level. For the established clinical large-scale database of 1051 white-light endoscopic images, ten-fold cross-validation is used in method validation. Experiment results show that the proposed framework achieves classification with sensitivity of 0.9034, specificity of 0.9718, and accuracy of 0.9628, and the segmentation with sensitivity of 0.8018, specificity of 0.9655, and accuracy of 0.9462. All of these indicate that our method enables an efficient, accurate, and reliable esophageal lesion diagnosis in clinics.

Index Terms—Esophageal lesions; deep learning; dual-stream esophageal lesion classification; convolutional neural network (CNN)

I. Introduction

Accurate classification and segmentation of the esophageal lesions are effective tools to help clinicians make reliable diagnostic schemes intensively depending on the potential lesions image analysis on the basis of classification and segmentation for esophageal lesion. Accurate classification

* Corresponding authors:
chenyang.list@seu.edu.cn (Yang Chen) and help_007@126.com (Xiaopu He)
† These authors contributed equally to this work

for the esophageal lesions is important because it reveals the esophageal lesion statuses which can further determine the prognosis of patients with esophageal lesions (Hu, Hu et al. 2010). In advanced stages, the five-year survival rate of esophageal cancer is 20.9% while greater than 85% in the early stages (Janurova and Bris 2014). (2) Accurate segmentation for the esophageal lesions can provide the annotation information of esophageal lesion regions and the elaborate feature analysis of lesions sizes, shapes, and colors. Classification and segmentation for esophageal lesions are indispensable and complementary, which provides comprehensive information together for a thorough understanding of esophageal lesions in clinical studies.

Changes in esophagus mucosa are closely related to the stages of cancerous developments, which has important significance in classifying and segmenting esophageal lesions for the clinician (Wang and Sampliner 2008). Different stages of esophageal lesions produce physiological and visual variations in the esophagus mucosa. Typical white-light images with three types of esophageal lesions and one normal type are shown in Fig. 1. The normal type shows no apparent lesion areas in Fig. 1(a). The type Inflammation is featured by red and white mucosa with strip shapes as shown in Fig. 1(b). For type Barrett, a clear boundary between normal areas and lesion areas appears in the epithelium as shown in Fig. 1(c). Esophageal cancer refers to a malignant lesion formed by abnormal proliferation in the epithelium of esophageal, which has irregular mucosae, disordered, and missing blood vessels in esophageal images as shown in Fig. 1(d) (Zellerhoff, Lenze et al. 2016). These visual lesion differences from esophageal images make esophageal lesion classification own theoretical support.

However, due to individual lesion variations in shape, color, and texture, depending on the clinician experience for esophageal lesion detection still exists potential misjudgments and time-consuming problems. To overcome the aforementioned problems, automated esophageal lesion detection using computer vision methods can be used for lesion classification and segmentation (Domingues, Sampaio et al. 2019).

For the accurate and automated classification and segmentation from esophageal images, it is still an open and challenging task because: 1) Significant intra-lesion differences in shape, size, and color seriously hamper the classification and segmentation performance as shown in Fig. 2(a). 2) Inter-lesion similarities between two different lesion types easily fall into the same category as shown in Fig. 2(b). 3) Varying illumination and noise degrees such as specular reflection easily produce the negative impacts on the esophageal lesion classification and segmentation (Tanaka, Fujiwara et al. 2018).

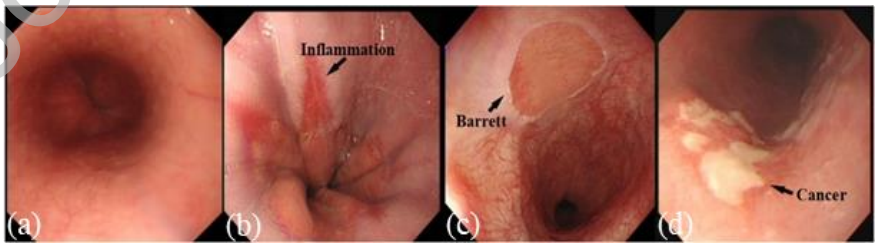


Fig. 1. Four types of white-light esophageal images: (a) The normal type represents there is no any lesion. (b) The type Inflammation is featured by red and white mucosa with strip shapes. (c) The type Barrett is featured by a clear boundary between normal areas and lesion areas appearing in the epithelium. (d) The type Cancer is characterized by the irregular mucosae, disordered and missing blood vessels.

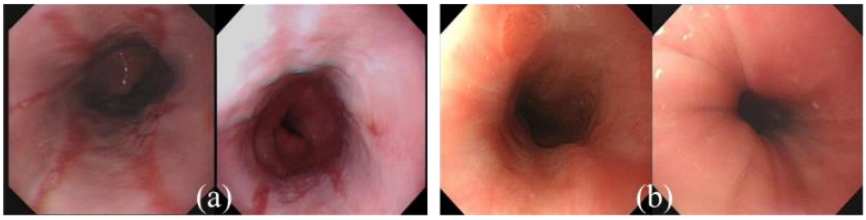


Fig. 2. Intra-lesion differences and Inter-lesion similarities make esophageal lesion classification and segmentation

Journal Pre-proof

challenging. (a) Intra-lesion differences in shape, size, and color for Inflammation lesion. (b) Inter-lesion similarities between the type Barrett and Normal.

1.1. Related Works

1. **Classification methods** aim to automatically predict esophageal lesion statuses for improving the diagnosis efficiency and precision. However, due to the individual differences from the same lesion types and visual similarities from different lesion types, accurate classification is challenging. Researchers have made some attempts including traditional methods based on prior knowledge and deep learning methods to tackle this challenge. Kandemir et al. (Kandemir, Feuchtinger et al. 2014) performed a study for the diagnosis of Barrett's esophagus presented in hematoxylin-eosin stained histopathological biopsies using multiple instance learning (MIL) and Support Vector Machines (SVMs). Souza Jr. et al. (Souza, Hook et al. 2017) classified lesions in Barrett esophagus using the SVMs classifier based on the descriptors of Speed-Up Robust Features (SURF). Li et al. (Li, Meng et al. 2009, Antony, McGuinness et al. 2016) presented a method based on Local Binary Patterns (LBPs) and curvelets for texture extraction aiming to distinguish ulcer and normal regions of esophagus using capsule endoscopy (CE). By considering the use of inaccurate prior knowledge and low-level handcrafted features (Antony, McGuinness et al. 2016), the supervised deep learning method tries to learn hierarchical and complicated representation of data features for esophageal lesion classification (Hong, Park et al. 2017, Horie, Yoshio et al. 2019). Hong, J et al. (Hong, Park et al. 2017) developed a computer-aided diagnosis (CAD) system using CNN architectures for classification among gastric metaplasia, intestinal metaplasia, and neoplasia. Yoshimasa et al. (Horie, Yoshio et al. 2019) employed the CNN to classify the esophageal cancer including squamous cell carcinoma and adenocarcinoma. Recently, CNNs have been proved effective methods for many medical imaging tasks, including feature recognition (Yan 2018), image analysis (Singh, Rote et al. 2018), and lesion detection (Tanaka, Fujiwara et al. 2018).

Discussion: All of these show great potential with the development of esophageal lesion classification. However, they do not make full use of the global and local lesion features and it causes the ambiguous representation for different esophageal type features. In this paper, we integrate dual-view contextual lesion information and complementarily and comprehensively extract global features as the basis of the holistic esophageal image information and local features as the details of the esophageal lesions. It effectively reduces inter-lesion similarities and intra-lesion differences for accurate and automatic four types of esophageal lesion classification.

2. **Segmentation methods** aim to annotate the ROIs (region of interests) of the esophageal lesions at the pixel level. It has undergone a great development (Van Der Sommen, Zinger et al. 2014, Georgakopoulos, Iakovidis et al. 2016, Mendel, Ebigbo et al. 2017). Van d S F et al. (Van Der Sommen, Zinger et al. 2014) computed local color and texture features based on the original images and Gabor-filtered images and utilized Support Vector Machines (SVMs) classifier for early esophageal cancers segmentation. With the development of deep learning, many researchers followed the trends and proposed deep learning methods for learning feature representations in the application of esophageal lesion segmentation. Georgakopoulos et al. (Georgakopoulos, Iakovidis et al. 2016) proposed a weakly-supervised learning technique based on CNNs, and it only used image-level semantic annotations for the training process instead of using annotations at the pixel level. Mendel et al. (Mendel, Ebigbo et al. 2017) employed the transfer learning based method to segment adenocarcinoma and Barrett's esophagus images and obtained the results with the sensitivity of 0.94 and specificity of 0.88.

Discussion: The above studies have made significant contributions to esophageal lesion segmentation. However, these algorithms still have potentials for improvement in the segmentation performance because they do not purposefully consider the differences of varying esophageal lesion

Journal Pre-proof

types in shape, color and size, which hampers capturing common lesion features. In this paper, to tackle this problem, we design the lesion-specific segmentation network to automatically annotate three lesion types (Inflammation, Barrett, Cancer) at pixel level.

1.2. Contributions

In this paper, we propose Esophageal Lesion Network (ELNet) based on deep CNNs to classify and segment the esophageal lesions with four interdependent functional parts: **Preprocessing module**, **Location module**, **Classification module**, and **Segmentation module**. (1) To normalize esophageal images, reduce obstruction of irrelevant information and tackle data imbalance problem, the **Preprocessing module** is used for normalization, specular reflection removal, and data augmentation from original esophageal images. (2) To highlight esophageal lesions, the **Location module** employs the Faster RNN for focusing on the ROIs of esophageal lesions. (3) To accurately predict esophageal lesion statuses, the **Classification module** is designed for classifying four esophageal lesion types. (4) To obtain accurate annotation at pixel level, the **Segmentation module** is employed to automatically segment three lesion types.

The main contributions of our work can be generalized as follows:

5 For the first time, we propose the ELNet for automatically and reliably comprehensive esophageal lesions classification and lesion-specific segmentation from clinically white-light esophageal images. It enables an efficient esophageal lesion detection to make suitable and repaid diagnostic schemes for clinicians.

6 A novel Dual-Stream Network (DSN) is proposed for esophageal lesion classification. DSN automatically integrates dual-view contextual lesion information using two CNN streams to complementarily extract the global and local features. It effectively improves the esophageal lesion classification performance to automatically predict the esophageal lesion statuses.

7 The lesion-specific esophageal lesion annotation with the Segmentation Network with Classification (SNC) strategy is proposed to reduce the intra-lesion differences for automatically segmenting three lesion types at pixel level.

8 A clinically large-scale database esophageal database is established for esophageal lesions classification and segmentation. This database includes 1051 white-light esophageal images, which consists of endoscopic images in four different lesion types. Each esophageal image in this database has a classification label and its corresponding segmentation annotation.

Experiment results show that the proposed ELNet for esophageal lesions achieves the classification results with sensitivity of 0.9397, specificity of 0.9825, and accuracy of 0.9771, and the segmentation results with sensitivity of 0.8018, specificity of 0.9655, and accuracy of 0.9462. All of these indicate that our method enables an efficient, accurate and reliable esophageal lesion diagnosis in clinics.

The remainder of this paper is organized as follows: In Section II, we present proposed ELNet for esophageal lesion classification and segmentation. In Section III, the implementation details about ELNet are reported. The experiment and evaluation are given to validate the performance of ELNet in Section IV. Finally, we conclude the proposed ELNet and discuss related future work in Section V.

II. Method

The proposed ELNet includes the following interactional functional parts: (1) the **Preprocessing module** performs the operations of normalization, spectral reflection removal, and data augmentation to normalize the esophageal images, reduce the irrelevant information obstruction, and tackle overfitting

problem; (2) the **Location module** employs Faster-RCNN to highlight the ROIs of the esophageal lesions; (3) the **Classification module** utilizes the proposed DSN consisting of Global and Local Streams to simultaneously extract the global and local features for four-class esophageal lesion classification (Normal, Inflammation, Barrett and Cancer); (4) the **Segmentation module** performs the automatic esophageal lesion annotation at pixel level using the proposed lesion-specific segmentation network. The main workflow is outlined in Fig. 3.

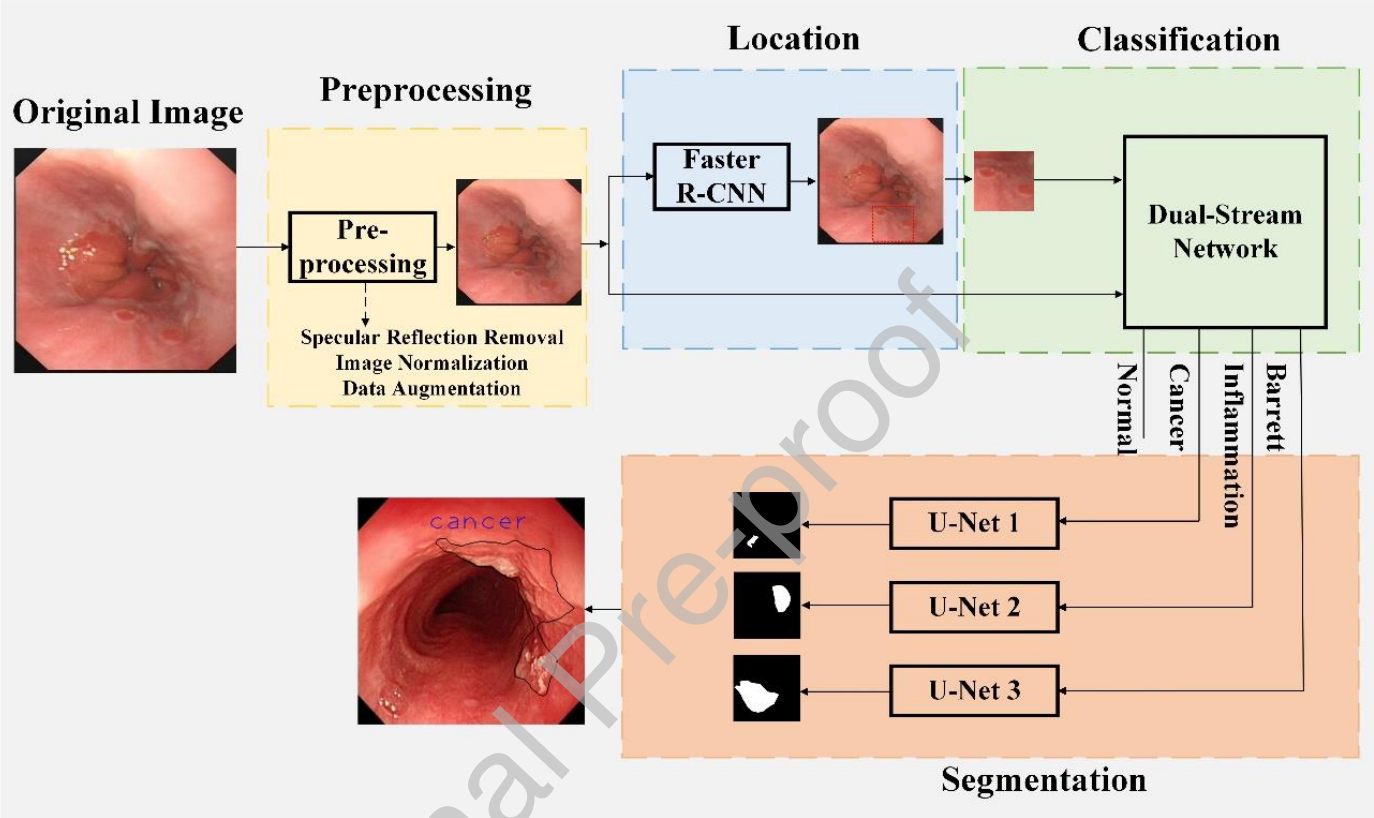


Fig. 3. Overview of the proposed ELNet for esophageal lesion classification and segmentation.

2.1. Preprocessing Module

The Preprocessing module consists of three steps: (1) the specular reflection removal is performed to reduce the random white spots in the endoscopic images; (2) the normalization operation is used to initialize esophageal images for dimension unification and computation complexity reduction; (3) the data augmentation is employed to tackle the overfitting problem.

2.1.1. Specular Reflection Removal

Specular reflections removal performs the specular reflection detection and correction to remove non-uniformity of the illumination (Tchoulack, Langlois et al. 2008). This non-uniformity of the illumination caused by the deviation of the light sources generates the specular reflection with white spots in endoscope images. It is implemented via two sub-tasks in a streaming mode:

(1) **Detection**: A bi-dimensional histogram decomposition is applied to detect the specular reflections:

$$m = \frac{1}{3}(r + g + b), \quad (1)$$

Journal Pre-proof

$$s=\begin{cases}\frac{1}{2}(2r-g-b)=\frac{3}{2}(r-m) & if\ (b+r)>2g\\ \frac{1}{2}(r+g-2b)=\frac{3}{2}(m-b) & if\ (b+r)\geq 2g\end{cases}, \quad (2)$$

where m is the pixel intensity, s is the saturation, and r, g and b represent the red, green and blue channels in images. Specular reflections can be detected via two threshold values m_{max} and s_{max} based on the bi-dimensional histogram. A pixel p will be a part of the specular region if it meets the following conditions:

$$\begin{cases} m_p \geq \frac{1}{2} m_{max} \\ s_p \leq \frac{1}{3} s_{max} \end{cases}, \quad (3)$$

where m_{max} and s_{max} are the maximum intensities of m_p and s_p among all the pixels in an image respectively. Related parameters are obtained by experiment with a large quantity of esophageal images.

(2) **Correction**: The Navier-Stokes method (Bertalmio, Bertozzi et al. 2001) performs the linear correction. The corrected images are obtained by replacing the specular reflection points with the average neighboring pixel values. The images before and after this spectral reflection removal are depicted in Fig. 4.

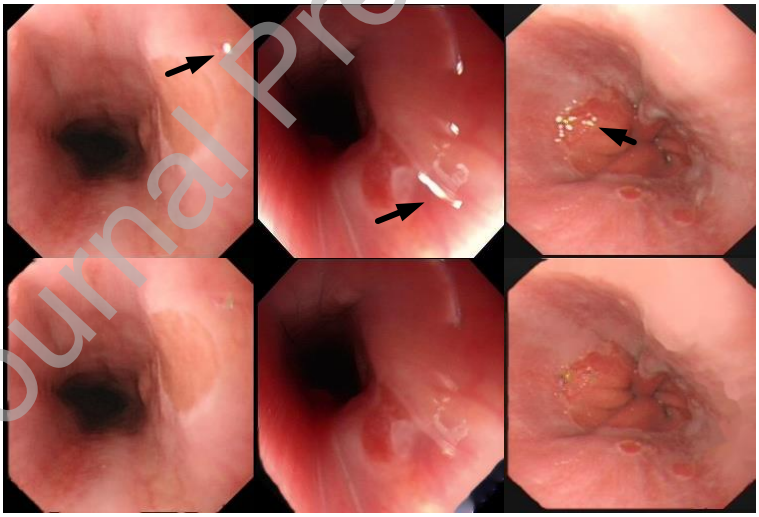


Fig. 4. The spectral reflection removal for original esophageal images (top) and corresponding images after spectral reflection removal (bottom). The non-uniformity of the illumination caused by the deviation of the light sources generates the specular reflection with white spots in endoscope images. The arrows in the original esophageal images point to these white spots.

2.1.2. Image Normalization

Image normalization performs the image standardization and lowering computational complexity. The esophageal images are normalized into 512×512 pixels as the input of the Global Stream and the lesion patches as the input of the Local Stream are resized into 64×64 pixels in the Classification module.

The pixel values of three channels for esophageal images are normalized into the range of (0, 1). The normalization formula is defined by Eqs (4):

$$y = \frac{x - MinValue}{MaxValue - MinValue}, \quad (4)$$

Journal Pre-proof

where x is the input pixel value of an esophageal image, $MinValue$ and $MaxValue$ is the minimum and maximum pixel value of this esophageal image, y is the pixel value output after normalization.

2.1.3. Data Augmentation

Data augmentation tackles the over-fitting problem. It includes the translation to simulate the left and right position change of gastroscopes, the rescaling to simulate the gastroscope stretch, the rotation to simulate the rotation movement of gastroscope, and the flipping to simulate the mirroring of gastroscope in our experiment. A summary of the transformations with the parameters is given in TABLE 1.

TABLE 1.
Data Augmentation Parameters.

Transformation Type	Description
Rotation	Randomly rotate an angle of $0^{\circ} - 360^{\circ}$
Flipping	0 (without flipping) or 1(with flipping)
Rescaling	Randomly with scale factor between 1/1.6 and 1.6
Translation	Randomly with shift between −10 and 10 pixels

2.2. Location Module

The Location module performs the lesion location and generates corresponding lesion patches for four esophageal image types. In this paper, Faster RCNN (Everingham, Van Gool et al. 2007) including Region Proposal Networks (RPN) and Fast R-CNN (Singh, Rote et al. 2018) is utilized to locate the ROIs of lesions. The RPN predicts the possibility of an anchor being background or foreground and refines the anchor. In this paper, there are the anchors with the fixed 1:1 length-width ratio to avoid distortion of lesion shapes when resizing the images. The VGG16 network (Qassim, Verma et al. 2018) is employed to extract spatial features. Fast R-CNN performs the final classification and regression to locate the ROIs of lesions.

For the three esophageal lesion types, the length/width values and coordinates of bounding boxes are obtained, and the corresponding location lesion areas are clipped as lesion patches. For the normal type, the image patches are captured by randomly clipping the normal type of esophageal images. Four types of esophageal image patches are resized into 64×64 pixels to be the input of the Local Stream of the DSN for esophageal image classification. This normalization can effectively reduce the intra-class difference from the physical size of esophageal lesion, which conforms to real-time clinical diagnostic environment. Fig. 5 shows the four types of esophageal images and corresponding image patches. Compared with the holistic esophageal images, clipped lesion patches focuses on the local and detailed features such as colors and textures, which provide the necessity for classification of four esophageal lesion types to reduce the inter-lesion similarities and intra-lesion differences.

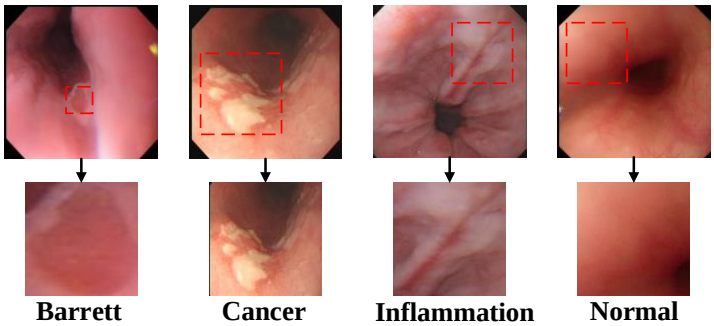


Fig. 5. Four types of esophageal images and the corresponding image patches. The red blankets in the top figure are the lesion areas located by Faster RCNN and the bottem figure show the clipped image patches.

2.3. Dual-Stream Network for Classification Module

Journal Pre-proof

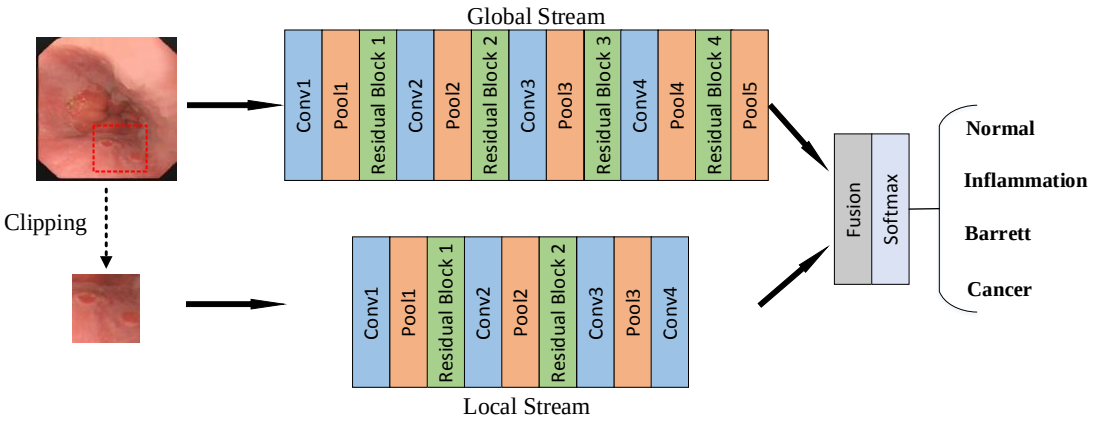


Fig. 6. The structure of the proposed Dual-Stream Network (DSN).

The Classification module performs the classification for four esophageal image types including Normal, Inflammation, Barrett, and Cancer. To automatically integrate dual-view contextual lesion information and accurately classify four esophageal lesion types, the DSN is designed, and it consists of two complementary streams – the **Global Stream** and the **Local Stream**. The Global Stream extracts global features via inputting the holistic esophageal images to focus on the holistic esophageal image information reflecting the contrast between lesion regions and background. The Local Stream extracts local lesion features related to information about textures, shapes, and colors of lesions via inputting four types of lesion patches generated by the Location module. The structure of the DSN for esophageal lesions is depicted in Fig. 6.

The detailed configuration of the proposed DSN is shown in TABLE 2. Given the input data scale and image size, the Global Stream has 21 layers including 16 convolution layers and 5 pooling layers. The stride of convolution layer is set to 1 to capture the lesion features besides the Conv 1 with stride 2 to reduce the computation parameters in the Global Stream. The Local Stream is designed to have 13 layers including 10 convolution layers and 3 pooling layers. The Conv 4 is added in the Local Stream to keep the same output sizes with the Global Stream. Each convolution layer is followed by a batch normalization layer and a ReLU activation layer.

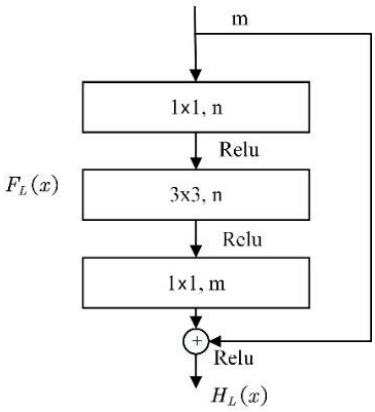


Fig. 7. A building block of residual network.

The Global Stream and the Local Stream are designed based on the ResNet structure proposed by Kaiming He et.al (He, Zhang et al. 2016). The ResNet network can well tackle degradation problem via shortcut connection, and Fig. 7 depicts the structure of the Residual block (Szegedy, Ioffe et al. 2017, Wu, Shen et al. 2019). For clarity, $H_L(x)$ denotes the transformation function of the L^{th} building block, and x is the input of the L^{th} building block. The desired output for the Residual block is set as $F_L(x)$. The Residual block explicitly forces the output to fit the residual mapping, i.e. the stacked nonlinear

layers are used to learn the following transformation:

$$F_L(x) = H_L(x) - x \quad \text{.....(5)}$$

therefore, the transformation for the L^{th} building block is:

$$H_L(x) = F_L(x) + x \quad \text{.....(6)}$$

The residual block consists of convolution layers with the kernel size of 1×1 and 3×3. The convolution layers with kernel size 1×1 are used to reduce channel number into n ($n<m$) and the convolution layers with kernel size 3×3 are employed for extracting spatial features and returning to the input channel number m. In this paper, four residual blocks are built in the Global Stream to extract the global features, and two residual blocks are built in the Local Stream to extract the local features.

The pooling layer is utilized for down-sampling to reduce computation complexity. From the detailed design of the proposed DSN in Table 2, we can see that the data size is downsampled from 512×512 pixels in Global Stream and 64×64 pixels in Local Stream to 8×8 pixels throughout the pooling layers together.

The concatenation fusion is used to fuse the output of the Global Stream and the Local Stream (Chen, Xie et al. 2018). For clarity, we define a fusion function $y = f_{cat}(x^a, x^b)$, two feature maps $x^a \in R^{H\times W\times D}$ and $x^b \in R^{H\times W\times D}$, and a fusion feature map $y \in R^{H'\times W'\times D'}$ (W, H, and D are the width, height and number of channels of feature maps). The function $y = f_{cat}(x^a, x^b)$ stacks the two features at the same location (i, j) across the feature channel d:

$$y_{i,j,d} = x^a_{i,j,d} \quad y_{i,j,D+d} = x^b_{i,j,d} \quad \text{.....(7)}$$

where $y \in R^{H\times W\times 2D}$. Then, the fused feature map has 1024 features with the size of 8×8 pixels. Finally, these features are flattened and inputted into the Softmax layer.

The Softmax layer performs normalization for feature maps into the range of (0, 1) so that the output vector $\hat{y}_m$ represents the probability of the m^{th} class. The operation for the Softmax layer can be written as:

$$\hat{y}_m = \frac{e^x}{\sum_{m=1}^{4} e^x} \quad \text{.....(8)}$$

where $\hat{y}_m$ is the output probability of the m^{th} class, x represents the input neurons of the upper layer.

We choose cross entropy loss as the objective function of DSN to accelerate training. The cross-entropy based loss function is given by (9):

$$loss = -mean(y \times log(\hat{y}_m) + (1-y) \times log(1-\hat{y}_m)) \quad \text{.....(9)}$$

where, y is the label vector and y_m is the predicted output vector of the proposed DSN.

TABLE 2
The detail configuration of proposed DSN for esophageal lesion classification.

Dual-Stream Network (DSN)							
Global Stream				**Local Stream**			
Layer		**Kernel Size, Channel Number**	**Output Size**	**Layer**		**Kernel Size, Channel Number**	**Output Size**
Data		-	512×512	Data			64×64
Conv 1		3×3, 64	256×256	Conv1		3×3,64	64×64
Pool 1		2×2, 64	128×128	Pool 1		2×2, 64	32×32
Residual	Conv 1	1×1, 32	128×128	Residual	Conv1	1×1, 32	32×32

Journal Pre-proof

Block-1	Conv 2	3×3, 32	128×128	Block-1	Conv2	3×3, 32	32×32
	Conv 3	1×1, 64	128×128		Conv3	1×1, 64	32×32
Conv2		3×3,128	128×128	Conv2		3×3,128	32×32
Pool 2		2×2, 128	64×64	Pool 2		2×2, 128	16×16
Residual Block-2	Conv 1	1×1, 64	64×64	Residual Block-2	Conv 1	1×1, 64	16×16
	Conv 2	3×3, 64	64×64		Conv 2	3×3, 64	16×16
	Conv 3	1×1,128	64×64		Conv 3	1×1,128	16×16
Conv3		3×3,256	64×64	Conv3		3×3,256	16×16
Pool 3		2×2, 256	32×32	Pool 3		2×2, 256	8×8
Residual Block-3	Conv 1	1×1, 128	32×32	Conv 4		1×1, 512	8×8
	Conv 2	3×3,128	32×32	-		-	-
	Conv 3	1×1,256	32×32	-		-	-
Conv 4		3×3, 512	32×32	-			
Pool 4		2×2, 512	16×16	-		-	-
Residual Block-4	Conv 1	1×1, 256	16×16	-		-	-
	Conv 2	3×3, 256	16×16	-		-	-
	Conv 3	1×1, 512	16×16	-		-	-
Pool 5		2×2, 512	8×8	-		-	-
Fusion		8×8, 1024					
Softmax		9 Neurons					

2.4. Segmentation Module

The Segmentation module performs the lesion-specific esophageal annotation automatically for three esophageal lesion types including Inflammation, Barrett, and Cancer at pixel level. Given the differences of shapes, sizes, colors, and textures for these three types of lesions, different segmentation strategies have strong effect for the segmentation results. To reduce these intra-lesion differences, we proposed lesion-specific segmentation strategies where three types of esophageal lesions are segmented respectively based on Classification module. It is referred as Segmentation Network with Classification (SNC). This strategy is more targeted for each lesion type and reduces the adverse impacts of inter-lesion. To demonstrate the superiority of this strategy, we compared with the most straightforward strategy which segments all the esophageal lesion images with no classification directly in the experiment part (Van Der Sommen, Zinger et al. 2014, Xue, Zhang et al. 2017). It is referred as Segmentation Network with No Classification (SNNC).

Our proposed lesion-specific segmentation network is designed based on the U-Net architecture with good performance in medical imaging segmentation (Ronneberger, Fischer et al. 2015). A typical U-Net architecture consists of the contracting path, symmetrical expanding path, and bottleneck layer.

The contracting path performs the lesion feature extraction, and it has four typical contraction blocks. The contraction block consists of two convolution layers with kernel sizes 3×3 and a max-pooling layer with kernel size 2×2 in this paper. The contraction path captures lesion-specific features and the image size is reduced from 512×512 pixels to 32×32 pixels. The expanding path performs two-class classification for each image pixel, and it consists of four typical expansion blocks. The contraction block contains one deconvolution layer with stride 2 to reduce computation parameters, one concatenation layer to fuse the features with corresponding layers of contracting path, and two convolution layers with kernel size 3×3 to integrate lesion features. The output vectors of the expanding path represent the probability values of the foreground and background. The bottleneck layer is designed from two convolution layers with kernel size 3×3 for feature integration. The binary cross-entropy loss function is employed in loss calculation.

III. Materials and Implementation Details

Journal Pre-proof

3.1. Material.

The clinical database containing standard 1051 white-light endoscopic images from 748 patients is obtained from the First Affiliated Hospital of Nanjing Medical University between July 2017 and July 2018. The inclusion criteria of this database is the selection of those images with available conventional white-light endoscopy and pathologic analysis. Those images with poor quality and images captured from patients undergoing surgical or endoscopic resection are excluded. All the included esophageal images contain the pixel-level lesion annotations manually marked by licensed physicians and four types of the esophageal lesion labels based on strict histological proof.

Ten-fold cross-validation is used. 80 percent of the esophageal images are used for training, and 10 percent of the images are used for testing. The rest of the images are used as the validation dataset to optimize the training parameters and improve the generalization capability of the proposed network. It is noted that there is no data overlapping between the training dataset, validation dataset, and test dataset. The detailed statistics of collected esophageal images can be seen in TABLE 3.

TABLE 3
Statistics distribution from esophageal image database.

	Normal	Inflammation	Cancer	Barrett
Train	203	345	100	207
Validation	25	43	12	25
Test	26	44	14	27
Total	254	412	126	259

3.2. Implementation details

The computer platform is configured as follows: CPU was Inter(R) Core(TM) i7-5930K 3.5GHz; GPU was NVIDIA 2080TI with 11G memory. All codes were written under Python 3.6, and we used Tensorflow r1.4 as the deep learning library. The CUDA edition used here was 10.0. During the training phase, the weight parameters are trained using mini-batch stochastic gradient descent with momentum (set to 0.9). The base learning rate is set to 10^{-4} and iteratively decreases until the loss stops decreasing. All the models were trained using 10000 iterations. For the **Location module**, the VGG-16 is fine-tuned based on the pre-trained ImageNet model. The training batch size is set to 8 for gradient descent. For the **Classification module**, the training batch size is set to 8 for gradient descent. In the testing phase, all the preprocessed test images of 512×512 pixels and 64×64 pixels are input into the DSN simultaneously. For the **Segmentation module**, the training batch size is set to 4 for the three U-Net segmentation networks. In the testing phase, three subsets (Cancer, Inflammation, and Barrett) from classification results are inputted into these three trained U-Net networks (U-Net 1, U-Net 2, and U-Net 3) respectively.

Evaluation criteria. Quantitative and qualitative results of the proposed method are given in this section. With the collected esophageal images, we measure two system properties: (1) **Classification performance**, i.e., how many of the esophageal images are classified correctly, (2) **Segmentation performance**, i.e., how well do the annotations made by the algorithm match those of the expert physicians. To further clarify the evaluation metrics, the classification and segmentation results are measured using sensitivity (*SENS*), specificity (*SPEC*), accuracy (*ACC*), and the Receiver Operating Characteristics (*ROC*). The sensitivity, specificity, and accuracy are defined by Eqs. (10) - (12):

$$SENS = \frac{TP}{TP + FN}, \quad (10)$$

Journal Pre-proof

$$SPEC = \frac{TN}{FP + TN}, \quad (11)$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}, \quad (12)$$

where:

- TP, True Positives: the number of positive samples (foreground pixels) is correctly classified.
- FP, False Positives: the number of negative samples (background pixels) is wrongly classified as positive.
- TN, True Negatives: the number of negative sam-ples (negative pixels) is correctly classified.
- FN, False Negatives: the number of positive sam-ples (foreground pixels) is wrongly classified as negative.

IV. Experiment and Results

The ELNet demonstrates high performance with classification with sensitivity of 0.9034, specificity of 0.9718 and accuracy of 0.9628, and the segmentation with sensitivity of 0.8018, specificity of 0.9655 and accuracy of 0.9462. It indicates the effectiveness of our method in esophageal lesion classification and segmentation.

4.1. DSN for esophageal lesion classification

The experimental result shows the DSN accurately classifies four esophageal types and effectively integrates the advantages of the Global Stream and the Local Stream in TABLE 4. The DSN achieves an overall classification with sensitivity of 0.9034, specificity of 0.9718, and accuracy of 0.9628. Compared with the Global Stream and Local Stream, the DSN outperforms its subnetworks including the Global Stream in the term of sensitivity (6.04% improvement), specificity (3.09% improvement), accuracy (4.35% improvement) and the Local Network in the term of sensitivity (14.99% improvement), specificity (6.49% improvement) and accuracy (9.00% improvement). As shown in Fig. 8, The DSN achieves superior performance (AUC: 0.994) over the sole Global Network (AUC: 0.976) and sole Local Network (AUC: 0.971). These improvement results demonstrate that the DSN performs reasonable integration of these two streams to improve the classification performance and explains that the Global Stream is the basis of DSN, and the Local Stream is a powerful supplement for the Global Stream. These two streams working complementarily and comprehensively contribute to the DSN for esophageal lesion classification.

TABLE 4
The Comparative sensitivity, specificity, accuracy results of the proposed DSN and subnetworks (the Global Stream and the Local Stream).

	SPEC	SENS	ACC
Global Stream	0.9516	0.8793	0.9336
Local Stream	0.9176	0.7898	0.8871
DSN	**0.9825**	**0.9397**	**0.9771**

Journal Pre-proof

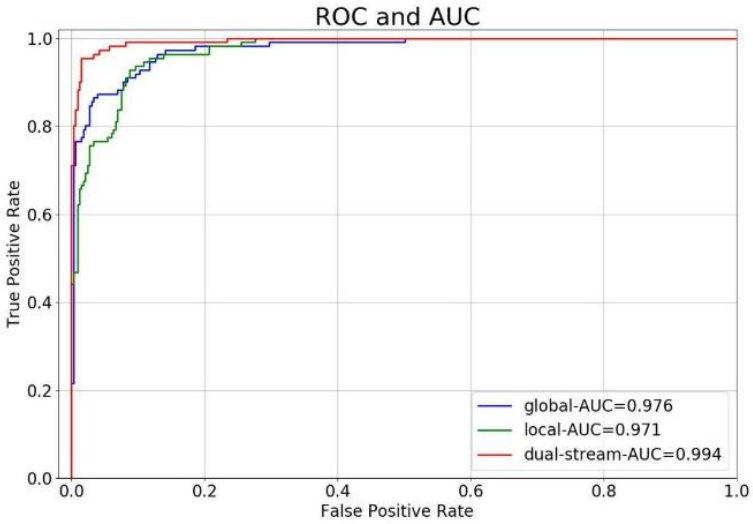
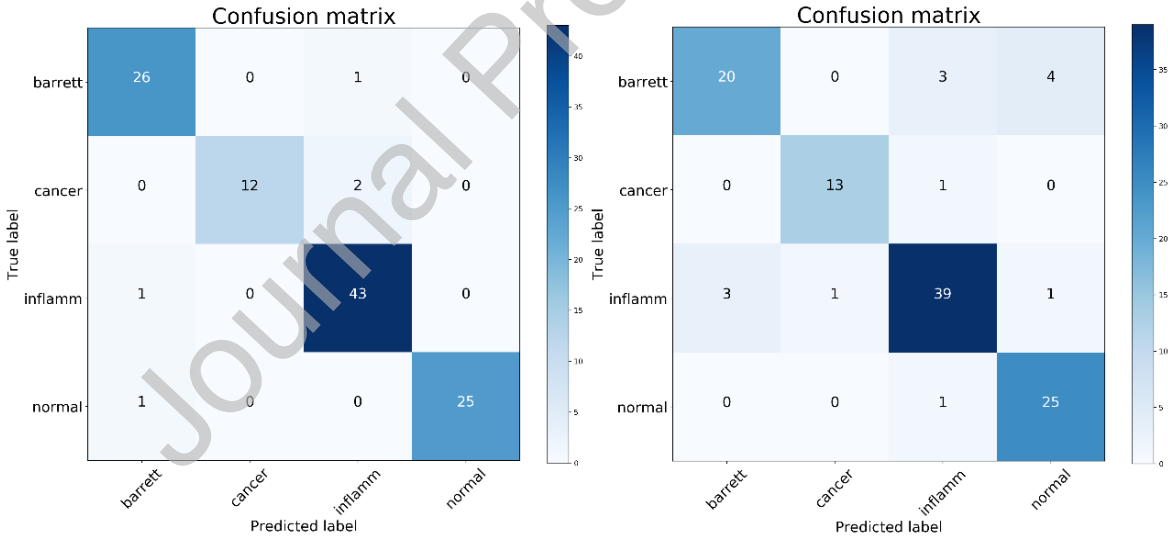


Fig. 8. ROC curves and AUC values of the DSN (Red), the Global Network (Blue) and the Local Network (Green).

As shown in Fig. 9, the DSN reduces the confusion degree in comparison with the Global Stream. The confusion matrix is a quantitative graph used to reflect the performance of a classification method on a testing set (Zhang, Shao et al. 2006). The diagonal values represent the number of correct classification for each class, and the others represent the confusion number between every two classes. The DSN increases the correct classification number in Inflammation (4 increased) and Barrett (6 increased). The type Cancer will get more excellent performance if more data can be available.



(a) (b)

Fig. 9. Confusion matrixes of (a) the proposed DSN and (b) the Global Stream in the esophageal image database.

To help understand the features extracted by the COVIDNet, we compute the class activation map (CAMs) from the Dual-Stream Network (DSN) (Selvaraju, Cogswell et al. 2017). The visualization results of the proposed DSN on the three esophageal lesion types are shown in Fig. 10. The closer to red in the heatmaps, the stronger activation in the original image, which indicates that information from that area contributes more to the final decision. As it can be seen from Fig. 10, the proposed DSN efficiently extracts esophageal lesion features, suppresses the irrelevant background information, and achieves excellent classification performance.

Journal Pre-proof

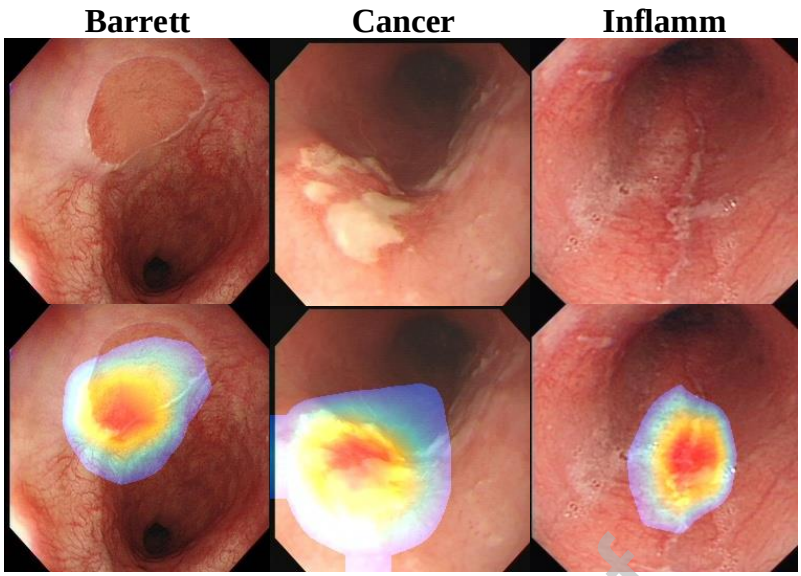


Fig.10. Visualization results of three types of esophageal lesion images (top) and the corresponding heatmaps (bottom) of the proposed DSN. The pixel areas related to esophageal lesions can be accurately highlighted by the proposed DSN for three types of esophageal lesion images in our database.

4.2. Classification comparison with state-of-the-art methods

As shown in TABLE 5., the DSN outperforms the other well-known classification methods including LBP+SVMs (Van Der Sommen, Zinger et al. 2013), HOG+SVMs (Kothari, Wu et al. 2016), and VGG-19 (Szegedy, Ioffe et al. 2017), ResNet-50 (He, Zhang et al. 2016) in all the terms of metrics SENS, SPEC, and ACC. LBP+SVMs and HOG+SVMs are traditional methods based on prior knowledge to extract hand-crafted features and using SVMs for classification. VGG-19 and ResNet-50 are deep learning based methods consisting of multiple convolutional and pooling layers alternately. After conducting a sufficient amount of experiments about parameter adjustment depending on our esophageal lesion database, the best results were kept to achieve the best classification performance. Overall, by comparison, it is observed in Table 5 that the proposed DSN achieves better performance than traditional method including LBP+SVM in the term of sensitivity (35.51% improvement), specificity (13.95% improvement), accuracy (48.39% improvement), and HOG+SVM in the term of sensitivity (24.93% improved), specificity (10.29% improved), accuracy (37.21% improvement). Among the deep learning based methods, our proposed DSN obtains better results compared with VGG-19 in the term of sensitivity (16.28% improvement), specificity (6.60% improvement), accuracy (9.07% improvement) and ResNet-50 in the term of sensitivity (5.99% improvement), specificity (2.83% improvement), and accuracy (4.25% improvement).

TABLE 5.
The Comparative sensitivity, specificity, accuracy results of proposed DSN and other well-known classification methods.

	SPEC	SENS	ACC
LBP+SVMs (Van Der Sommen, Zinger et al. 2013)	0.8430	0.5846	0.4932
HOG+SVMs (Kothari, Wu et al. 2016)	0.8796	0.6904	0.6050
VGG-19 (Szegedy, Ioffe et al. 2017)	0.9165	0.7769	0.8864
ResNet-50 (He, Zhang et al. 2016)	0.9542	0.8798	0.9346
Dual-Stream Network (DSN)	**0.9825**	**0.9397**	**0.9771**

4.3. Lesion-specific segmentation of esophageal lesions

As shown in TABLE 6, the proposed lesion-specific esophageal lesion segmentation can accurately automatic annotate the lesion ROIs at pixel level. Our proposed method achieves excellent

segmentation performance in all three average metrics including sensitivity of 0.8018, specificity of 0.9655, and accuracy of 0.9462. Compared with SNNC strategy, the lesion-specific segmentation with SNC significantly increases the effectiveness and stability in the terms of accuracy (6.96% improvement), sensitivity (21.80% improvement), and specificity (6.87% improvement). Fig. qualitatively compares the segmentation performance of between the proposed SNC strategy and the SNNC strategy. The lesion-specific segmentation with SNC strategy has a good match with the ground truth made by the specialist. For the SNNC strategy, the segmentation results of type Cancer and Inflammation produce a relatively higher false positive at pixel level due to under-fitting in three esophageal lesion types. These increases of our method are resulted from the fact that the lesion-specific segmentation provides an independent and efficient network for every esophageal lesion type and it adapts to each lesion type and reduces false positive at pixel levels.

TABLE 6.
The quantitative segmentation results of the SNC strategy and the SNNC strategy for esophageal lesions. The results of SNNC strategy are present in the brackets.

	Inflammation	Cancer	Barrett	Average
ACC	**0.9282**	**0.9075**	**0.9915**	**0.9462**
	(0.8806)	(0.7676)	(0.9152)	(0.8766)
SENS	**0.6909**	0.8020	**0.9387**	**0.8018**
	(0.5824)	**(0.8455)**	(0.5315)	(0.5838)
SPEC	**0.9648**	**0.9337**	**0.9954**	**0.9655**
	(0.9095)	(0.7374)	(0.9462)	(0.8968)

4.4. Segmentation comparison with state-of-the-art methods

As shown in TABLE 7, our proposed lesion-specific segmentation network achieves higher segmentation performance in all three metrics compared to the existing methods including the FCN (Noh, Hong et al. 2015), SegNet (Badrinarayanan, Kendall et al. 2017), and U-Net- SNNC (Ronneberger, Fischer et al. 2015). After conducting a sufficient amount of experiments about parameter adjustment depending on our esophageal lesion database, the best results were kept to achieve the best segmentation performance. Overall, by comparison, our proposed method obtains better segmentation results in all three metric terms. The proposed method achieves better performance than FCN in the term of sensitivity (39.26% improvement), specificity (0.04% improvement), accuracy (4.05% improvement), and SegNet in the term of sensitivity (14.31% improved), specificity (1.41% improved), accuracy (2.70% improvement), and U-Net-SNNC in the terms of accuracy (6.96% improvement), sensitivity (21.80% improvement), and specificity (6.87% improvement). It indicates the proposed methods can completely annotate the lesion areas for three esophageal types under high specificity, sensitivity and accuracy.

TABLE 7.
The Comparative sensitivity, specificity, accuracy results of employed segmentation network and other well-known segmentation methods.

Methods	SPEC	SENS	ACC
FCN (Noh, Hong et al. 2015)	0.9651	0.4092	0.9057
SegNet (Badrinarayanan, Kendall et al. 2017)	0.9514	0.6587	0.9192
U-Net-SNNC (Ronneberger, Fischer et al. 2015)	0.8968	0.5838	0.8766
Proposed Method	**0.9655**	**0.8018**	**0.9462**

Journal Pre-proof

Journal Pre-proof

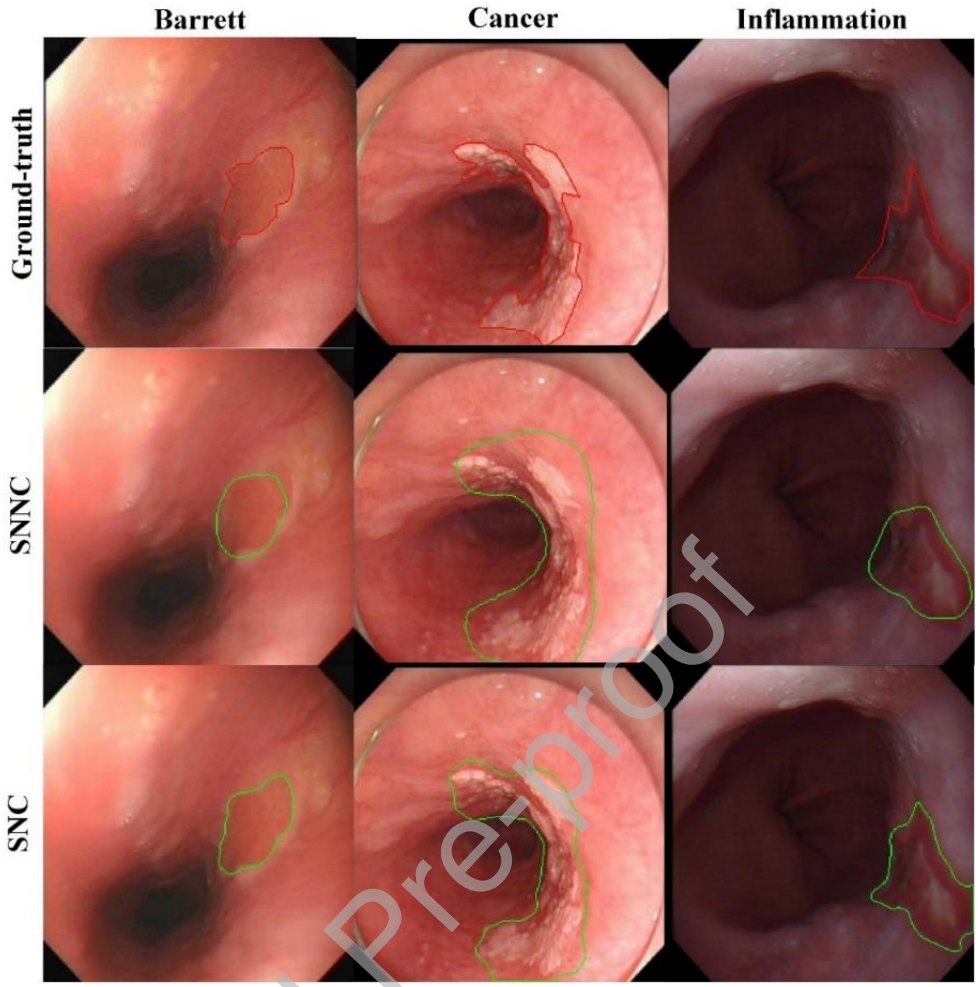


Fig. 11. Qualitative segmentation results of both the SNC and SNNC strategy for three types of esophageal lesions (Barrett, Cancer and Inflammation). The figures above show the ground-truth highlighted with a red line and the figures below show the predicted mask (the SNC and SNNC) highlighted with a green line. The three columns represent these three different types of esophageal lesions, respectively.

V. Conclusion and discussion

In this paper, we proposed the framework for the first time automatically achieves accurate classification and segmentation of esophageal lesions. The framework consists of four interdependent parts for esophageal lesion classification and segmentation: (1) the Preprocessing module is used for normalization, specular reflection removal, and data augmentation from original esophageal images to initialize the data; (2) the Location module employs the Faster RNN to focus on the ROIs of esophageal lesions; (3) In the Classification module, DSN integrates dual-view contextual lesion information to simultaneously extract global features and local features for esophageal lesion classification (Normal, Inflammation, Barrett, and Cancer); (4) the Lesion-specific annotation is proposed to segment three lesion types (Inflammation, Barrett, and Cancer) for automatic and accurate annotation at pixel level in segmentation module.

For the whole experiment, a clinical database of 1051 white-light endoscopic images is built. Ten-fold cross-validation is utilized in the method validation. Experiment results have proven that our proposed framework reaches high classification and segmentation results in all three metrics and maintains excellent internal consistency with the ground truth, indicating its great potential in clinical esophageal lesion evaluation.

Future work for the proposed method includes: (1) More elaborate esophageal lesion classification can be required. For example, the Cancer type of esophageal lesion can be further divided into epidermoid

carcinoma and adenocarcinoma (Asan and Nature 2017). It is beneficial for estimating the esophageal lesion statuses and making suitable diagnostic schemes; (2) a semi-supervised CNN-based method for esophageal lesions is required due to lack of the classification labels and annotations when larger training databases in clinics(Ge, Yang et al. 2019, Ge, Yang et al. 2019, Yin, Zhao et al. 2019).

Acknowledgment

This research was supported in part by the State's Key Project of Research and Development Plan under Grant 2017YFA0104302, Grant 2017YFC0109202 and 2017YFC0107900, in part by the National Natural Science Foundation under Grant 61801003, 61871117and 81471752, in part by the China Scholarship Council under NO. 201906090145.

REFERENCES

Antony, J., K. McGuinness, N. E. O'Connor and K. Moran (2016). Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE.

Asan, U. and C. G. A. R. N. J. Nature (2017). "Integrated genomic characterization of oesophageal carcinoma." **541**(7636): 169.

Badrinarayanan, V., A. Kendall, R. J. I. t. o. p. a. Cipolla and m. intelligence (2017). "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." **39**(12): 2481-2495.

Bertalmio, M., A. L. Bertozzi and G. Sapiro (2001). Navier-stokes, fluid dynamics, and image and video inpainting. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, IEEE.

Chen, Y., H. Xie and H. J. I. C. V. Shin (2018). "Multi-layer fusion techniques using a CNN for multispectral pedestrian detection." **12**(8): 1179-1187.

Domingues, I., I. L. Sampaio, H. Duarte, J. A. Santos and P. H. J. I. A. Abreu (2019). "Computer vision in esophageal cancer: a literature review." **7**: 103080-103094.

Everingham, M., L. Van Gool, C. K. Williams, J. Winn and A. Zisserman (2007). "The PASCAL visual object classes challenge 2007 (VOC2007) results."

Ge, R., G. Yang, Y. Chen, L. Luo, C. Feng, H. Ma, J. Ren and S. Li (2019). "K-net: Integrate left ventricle segmentation and direct quantification of paired echo sequence." IEEE transactions on medical imaging **39**(5): 1690-1702.

Ge, R., G. Yang, Y. Chen, L. Luo, C. Feng, H. Zhang and S. Li (2019). "PV-LVNet: Direct left ventricle multitype indices estimation from 2D echocardiograms of paired apical views with deep neural networks." Medical image analysis **58**: 101554.

Georgakopoulos, S. V., D. K. Iakovidis, M. Vasilakakis, V. P. Plagianakos and A. Koulaouzidis (2016). Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions. 2016 IEEE international conference on imaging systems and techniques (IST), IEEE.

He, K., X. Zhang, S. Ren and J. Sun (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition.

Hong, J., B.-y. Park and H. Park (2017). Convolutional neural network classifier for distinguishing Barrett's esophagus and neoplasia endomicroscopy images. 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE.

Horie, Y., T. Yoshio, K. Aoyama, S. Yoshimizu, Y. Horiuchi, A. Ishiyama, T. Hirasawa, T. Tsuchida, T. Ozawa and S. J. G. e. Ishihara (2019). "Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks." **89**(1): 25-32.

Hu, Y., C. Hu, H. Zhang, Y. Ping and L.-Q. Chen (2010). "How does the number of resected lymph

Journal Pre-proof

nodes influence TNM staging and prognosis for esophageal carcinoma?" Annals of surgical oncology **17**(3): 784-790.

Janurova, K. and R. Bris (2014). "A nonparametric approach to medical survival data: Uncertainty in the context of risk in mortality analysis." Reliability Engineering & System Safety **125**: 145-152.

Kandemir, M., A. Feuchtinger, A. Walch and F. A. Hamprecht (2014). Digital pathology: Multiple instance learning can detect Barrett's cancer. 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), IEEE.

Kothari, S., H. Wu, L. Tong, K. E. Woods and M. D. Wang (2016). Automated risk prediction for esophageal optical endomicroscopic images. 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), IEEE.

Li, B., M. Q.-H. J. I. Meng and V. computing (2009). "Texture analysis for ulcer detection in capsule endoscopy images." **27**(9): 1336-1342.

Mendel, R., A. Ebigbo, A. Probst, H. Messmann and C. Palm (2017). Barrett's esophagus analysis using convolutional neural networks. Bildverarbeitung für die Medizin 2017, Springer**:** 80-85.

Noh, H., S. Hong and B. Han (2015). Learning deconvolution network for semantic segmentation. Proceedings of the IEEE international conference on computer vision.

Qassim, H., A. Verma and D. Feinzimer (2018). Compressed residual-VGG16 CNN model for big data places image recognition. 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), IEEE.

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

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision.

Singh, H., S. Rote, A. Jada, E. D. Bander, G. J. Almodovar-Mercado, W. I. Essayed, R. Härtl, V. K. Anand, T. H. Schwartz and J. P. J. J. o. n. Greenfield (2018). "Endoscopic endonasal odontoid resection with real-time intraoperative image-guided computed tomography: report of 4 cases." **128**(5): 1486-1491.

Souza, L., C. Hook, J. P. Papa and C. Palm (2017). Barrett's esophagus analysis using SURF features. Bildverarbeitung für die Medizin 2017, Springer**:** 141-146.

Szegedy, C., S. Ioffe, V. Vanhoucke and A. A. Alemi (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence.

Tanaka, K., M. Fujiwara and H. J. G. Toyoda (2018). "An unlikely lesion to be identified in the cervical esophagus." **155**(3): 610-612.

Tchoulack, S., J. P. Langlois and F. Cheriet (2008). A video stream processor for real-time detection and correction of specular reflections in endoscopic images. 2008 Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference, IEEE.

Van Der Sommen, F., S. Zinger and E. J. Schoon (2013). Computer-aided detection of early cancer in the esophagus using HD endoscopy images. Medical Imaging 2013: Computer-Aided Diagnosis, International Society for Optics and Photonics.

Van Der Sommen, F., S. Zinger, E. J. Schoon and P. J. N. De With (2014). "Supportive automatic annotation of early esophageal cancer using local gabor and color features." **144**: 92-106.

Wang, K. K. and R. E. J. A. J. o. G. Sampliner (2008). "Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus." **103**(3): 788-797.

Wu, Z., C. Shen and A. J. P. R. Van Den Hengel (2019). "Wider or deeper: Revisiting the resnet model

Journal Pre-proof

for visual recognition." **90**: 119-133.

Xue, D.-X., R. Zhang, Y.-Y. Zhao, J.-M. Xu and Y.-L. Wang (2017). Fully convolutional networks with double-label for esophageal cancer image segmentation by self-transfer learning. Ninth International Conference on Digital Image Processing (ICDIP 2017), International Society for Optics and Photonics.

Yan, H. (2018). Computer Vision Applied in Medical Technology: The Comparison of Image Classification and Object Detection on Medical Images. 2018 International Symposium on Communication Engineering & Computer Science (CECS 2018), Atlantis Press.

Yin, X., Q. Zhao, J. Liu, W. Yang, J. Yang, G. Quan, Y. Chen, H. Shu, L. Luo and J.-L. Coatrieux (2019). "Domain progressive 3D residual convolution network to improve low-dose CT imaging." IEEE transactions on medical imaging **38**(12): 2903-2913.

Zellerhoff, S., F. Lenze, H. Ullerich, A. Bittner, K. Wasmer, J. Koebe, C. Pott, L. Eckardt, G. J. P. Moennig and C. Electrophysiology (2016). "Esophageal and Mediastinal Lesions Following Multielectrode Duty- Cycled Radiofrequency Pulmonary Vein Isolation: Simple Equals Safe?" **39**(4): 316-320.

Zhang, P., J. Shao, J. Han, Z. Liu and Y. Yan (2006). Keyword spotting based on phoneme confusion matrix. Proc. of ISCSLP.

Journal Pre-proof

Graphical abstract

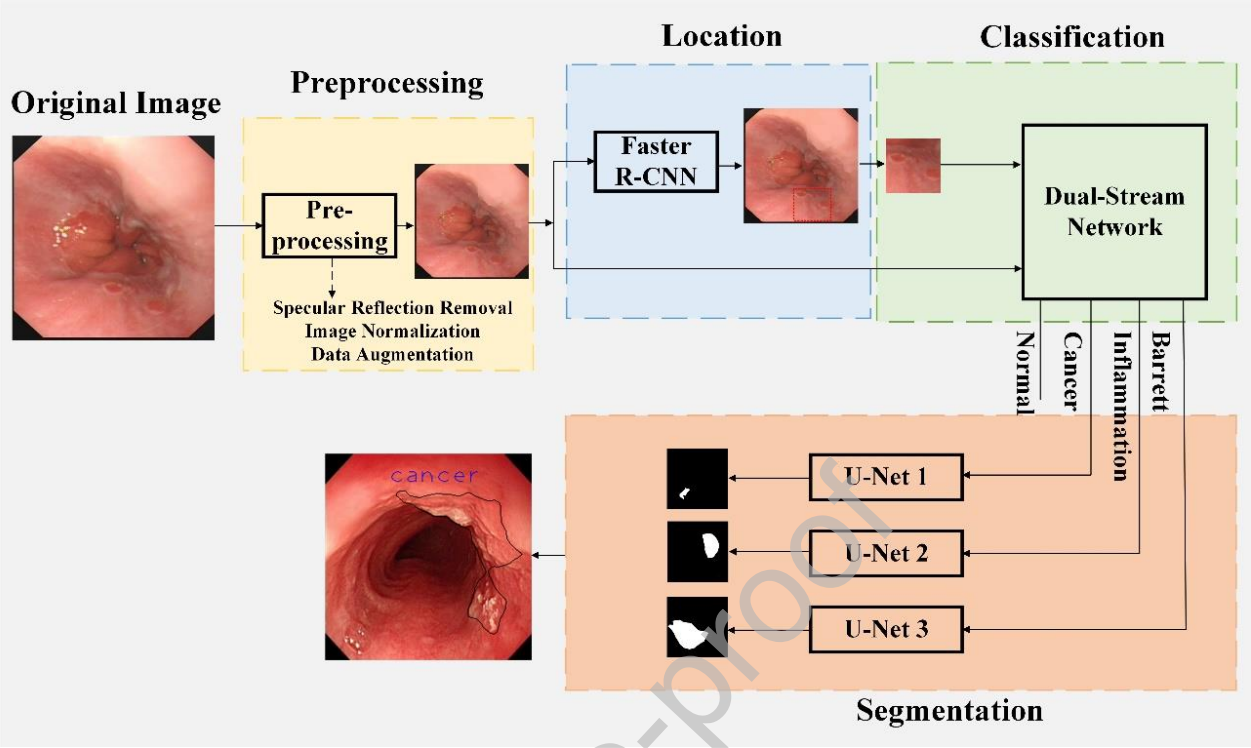


Fig. 3. Overview of the proposed ELNet for esophageal lesion classification and segmentation.

In this paper, we propose Esophageal Lesion Network (ELNet) based on deep CNNs to classify and segment the esophageal lesions with four interdependent functional parts: **Preprocessing module**, **Location module**, **Classification module** and **Segmentation module**. (1) To normalize esophageal images, reduce obstruction of irrelevant information and tackle data imbalance problem, **Preprocessing module** is used for normalization, specular reflection removal, and data augmentation from original esophageal images. (2) To highlight esophageal lesions, **Location module** employs the Faster RNN for focusing on the ROIs of esophageal lesions. (3) To accurately estimate esophageal lesion statuses and tackle the challenges of intra-lesion differences and inter-lesion similarities, **Classification module** is designed for classifying four esophageal lesion types (Normal, Inflammation, Barrett, and Cancer). (4) To obtain accurate annotation at pixel level, **Segmentation module** is employed to automatically segment three lesion types (Inflammation, Barrett, and Cancer).

Journal Pre-proof