Dual-Stage U-Shape Convolutional Network for Esophageal Tissue Segmentation in OCT Images

MENG GAN AND CONG WANG
Jiangsu Key Laboratory of Medical Optics, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
Corresponding author: Cong Wang (wang0112@sibet.ac.cn)

This work was supported in part by the Jiangsu Planned Projects for Postdoctoral Research Funds of China under Grant 2018K007A and Grant 2018K044C, and in part by the Jiangsu Province Science Fund under Grant BK20200216.

ABSTRACT Automatic segmentation is the crucial step for esophageal optical coherence tomography (OCT) image processing, which is able to highlight diagnosis-related tissue layers and provide characteristics such as shape and thickness for esophageal disease diagnosis. This study proposes a dual-stage framework using a specifically designed encoder-decoder network configuration for accurate and reliable esophageal layer segmentation, which is named as the dual-stage U-shape convolutional network (D-UCN). The proposed approach utilized one UCN to locate the target tissue region, which is followed by another UCN with similar architecture to achieve the final segmentation. In this way, the proposed strategy effectively solves the problems encountered in our previous studies, such as disturbance from neighboring diagnostically unrelated tissues, probe protection sheaths from the imaging equipment and the inevitable speckle noise. Experimental results on esophageal OCT B-scans from C57BL mice demonstrated that the proposed dual-stage framework achieved performance comparable to manual segmentation. The effectiveness and advantages of the dual-stage strategy are also confirmed in comparison with graph theory dynamic program (GTDP) and U-Net.

INDEX TERMS Esophageal layer segmentation, optical coherence tomography, fully convolutional network, medical image analysis.

I. INTRODUCTION
Esophageal diseases are receiving widespread attention due to their increasing incidence [1]. At present, the clinical examination of esophageal diseases by endoscopy is expensive and inconvenient [2], [3]. Optical coherence tomography (OCT) is a relatively new imaging technique [4], [5] that can avoid problems of traditional endoscopy. By combining fiber-optic flexible endoscopes, OCT devices are able to enter the upper gastrointestinal tract [5], [6] to image the microscopic structure of the esophagus, thereby helping diagnose a variety of esophageal diseases, such as Barrett’s esophagus (BE) [7], [8], eosinophilic esophagitis (EoE) [9], and dysplasia [10]. However, this diagnosis is laborious because it relies on analyzing the characteristics of esophageal tissue to accurately interpret a large number of images. Due to low-contrast images and disturbance (such as speckle noise and irrelevant structures), these tissues on OCT images are not easy to identify. As a result, automatic segmentation of esophageal OCT images and optical staining of target tissues will make it easy for tissue recognition, thus facilitating the diagnosis of esophageal diseases.

Developing an automatic segmentation system for esophageal OCT images has to address the following problems. Firstly, adjacent tissues that are irrelated to diagnosis (Fig. 1) should not be considered when segmenting. Secondly, structures such as probe-protected plastic sheath may be regarded as tissues by mistake when processing automatically by the computer. Finally, some blurred boundaries and the inherent speckle noise also have a negative impact on segmentation. Representative methods for automatic esophageal tissue layer segmentation can be summarized as follows. In 2016, Ughi et al. proposed an A-scan based method for esophageal lumen segmentation [11], but it can hardly be generalized for segmenting multiple internal tissue layers. Zhang et al. [12] employed the graph-based method [13]–[15] to segment five clinical-related tissue
layers, which is the first multi-layer esophageal tissue segmentation approach to our best knowledge. Based on the graph theory [13], our group improved Zhang’s method using modified Canny edge detection and realize more accurate segmentation [16]. The graph-based methods rely on prior boundary region information, and are sensitive to image quality and intensity variance, which may lead to failure in some cases. In 2019, our group proposed a more robust intelligent segmentation system based on the sparse Bayesian classifier using wavelet coefficients as features [17]. However, the feature extraction process for traditional machine learning requires considerable domain knowledge. The newly developed deep learning techniques have overcome these limitations and are able to automatically extract intrinsic features. Benefitting from such advantages, Li et al. proposed a U-Net based framework for an end-to-end esophageal layer segmentation [18].

Existing studies about deep learning methods applied to OCT images can be divided into two categories. The first category is boundary classification. Fang et al. segmented nine tissue layers in retinal OCT images by combining the convolutional neural network with graph search [19], and a similar method is adopted in Hamewood’s work [20]. Kugelman et al. identified the retinal boundaries using recurrent neural networks and graph search [21]. Such methods are implemented by classifying overlapped patches, which suffer from large redundancy and result in more time for inference [21]. The second category is pixel-wise segmentation. A typical strategy for methods of this kind is to perform pixel-wise label prediction using the fully convolutional network (FCN) [22]–[25]. FCN takes advantage of the highly convolved feature leaning ability of convolutional networks and utilizes the transposed convolution layer to magnify a feature map to the input size [26], [27]. Stemming from FCN, Ronneberger et al. proposed a more elegant U-shape network architecture called U-Net, which was designed for biomedical image segmentation with limited training images [28]. Inspired by these works, Roy et al. proposed a ReLayNet similar to U-Net for fluid segmentation in macular OCT image [29]. Devalla et al. designed the DRUNET for optic nerve head tissue segmentation in OCT image [30]. Venhuizen et al. implemented retinal thickness measurement and intraretinal cystoid fluid quantification based on U-shape FCN architecture [31]. Researchers have demonstrated that the second strategy possesses state-of-the-art performance on pixel-wise labeling [22], [23], [32] in biomedical image segmentation with limited training set. As a result, this strategy is also employed in the current work.

This study proposes a dual-stage U-shape convolution network (D-UCN) framework to segment esophagus in OCT images. Our main contributions can be summarized as follows:

- Propose a dual-stage framework D-UCN for esophagus segmentation. The network extracts the clinically interested region and removes non-significant information at the first stage. Then the obtained knowledge is used in the second stage to assist multiple esophageal layer identification.
- Design the UCN architecture which incorporates convolution block, residual block and mixed pooling methods to improve the classification performance.

The rest of this study is organized as follows. Section II describes the details of the proposed D-UCN framework for esophageal layer segmentation. Section III presents experimental settings and segmentation results on esophageal OCT images, including details of the dataset, comparisons with GTDP and U-Net. Discussions and conclusions are given in Sections IV and V, respectively.

II. METHODS

A. PROBLEM STATEMENT

Given an esophageal OCT image, the task is to assign each pixel to a particular label representing a certain tissue. A typical esophageal OCT image from the mouse is shown in Fig. 1(a). The target tissue layers marked in the images are the epithelium stratum corneum (SC), epithelium (EP), lamina propria and muscularis mucosae (LP & MM) and submucosa (SM), which are labeled from “1” to “4”, respectively. The remaining part of the image is labeled by “0” as displayed in Fig. 1(b).

![FIGURE 1. Demonstration of (a) a typical esophageal OCT image for mouse and (b) the corresponding manual segmentation result.](image-url)
network architecture, UCN-I and UCN-II, as shown in Fig. 3. The two networks adopted in this structure play different roles in our dual-stage learning scheme: UCN-I aims to define the region containing the layer of interest, while UCN-II intends to use the output of UCN-I as a constraint to segment the esophagus. Our framework is capable of adaptively addressing various OCT imaging problems and allows the introduction of a priori knowledge of target tissue areas to improve segmentation performance. More details for each network and the corresponding training procedure will be explained in the following sections.

1) UCN-I FOR ESOPHAGEAL TARGET TISSUE AREA IDENTIFICATION

The target tissue area accounts for a relatively small percentage of the entire OCT image. In addition, artifacts like plastic sheaths were also presented in the image as shown in Fig. 1(a). Those irrelevant parts were proved to have negative effects on segmentation performance to some extent [16], [17]. Restricting the learning process within a certain area that only contains layers we concerned is beneficial for determining which layer the pixel belongs to.

UCN-I is employed to extract the total tissue area labeled from “1” to “4” indicated in Fig. 1, which is a binary classification problem. After training with annotated images, the UCN-I is able to provide a probability for each pixel, which can be thresholded to generate a binary mask indicating the target tissue area.

2) UCN-II FOR ESOPHAGEAL LAYER SEGMENTATION

The binary mask obtained by UCN-I is employed as a priori information to assist the layer segmentation task of UCN-II. Specifically, we utilize the binary mask as a constraint to give pixels within the target region more importance and ignore structures outside the region, such as artifacts and background.

When segmenting a new OCT image, the Hadamard product of the image and the output of UCN-I is calculated to generate a new image with much less disturbance, which is subsequently sent to UCN-II for multi-layer recognition. In this case, more accurate segmentation is expected compared to segmenting the image directly.

3) NETWORK ARCHITECTURE

The network architecture used in the two phases is illustrated in Fig. 3. It is a fully convolutional U-shape network inspired by the U-Net, which is specifically designed for biomedical image segmentation [28]. The architecture is approximately symmetrical and can be divided into a contracting encoder and an expansive decoder, which are connected through concatenation layers. The three main components, including the encoder, the decoder and the classification layer, are detailed as follows.

The encoder consists of two convolutional blocks and one residual block as shown in Fig. 4. Both the convolution block and the residual block contain two $3 \times 3$ convolutional layers. The identity shortcut of the residual block is performed by a $1 \times 1$ convolution. For each block, a pooling layer is followed to downsample the received feature map. Max pooling and average pooling are two widely used pooling techniques that summarize the average presence and the most activated presence of a feature, respectively. Generally, max pooling enables the network to capture the most discriminative feature in a sub-region of kernel size and is often used as the default in CNNs, while average pooling is more robust than max pooling in the noisy case. Since the endoscopic OCT images are inevitably contaminated by the speckle noise, we combined these two pooling techniques in our network and make them work at different feature scales for better performance.
performance. As shown in Fig. 3, except for the first encoding block followed by a max pooling layer, each encoding block is followed by an average pooling layer. The max pooling layer set at the beginning is designed to distinguish the most informative features from the original resolution and ignoring other non-essential parts. The use of average pooling subsequently is to improve the robustness of the architecture and avoid information loss caused by max pooling.

In all encoding blocks, the convolutional layers are followed by a batch normalization layer and a PReLu activation layer. The employment of the batch normalization layer is to compensate for the covariate shifts and helps to achieve a successful training [29]. The PReLu activation is chosen because it can introduce non-linearity in the training and prevent gradient vanishment problem. Besides, the PReLu converges faster than ReLu [29]. For the residual block, the residual layers are batch normalized and the addition is followed by the PReLu. A dropout layer with a 0.5 dropout rate is applied at the end of the encoder to prevent overfitting.

The decoder is also composed of three blocks: two convolutional blocks and one residual block, which is similar to the encoder as shown in Fig. 3. The details of the two types of blocks have been described before. It aims to restore the feature map to the original resolution of the image and generate segmentation results. A $2 \times 2$ upsampling layer is employed using the saved indices from the pooling layers of the matched encoder block. The upsampling algorithm is bilinear interpolation which is simple and efficient [32], [33]. After upsampling, a skip connection is applied to concatenate the corresponding feature map from the encoder path with the upsampling feature maps, which intends to transfer local information to the global information from the decoder path while upsampling. The concatenation is necessary due to the loss of spatial information in convolution.

The classification layer is implemented by a $1 \times 1$ convolution layer with the number of filters equal to the number of categories. Then, an appropriate activation function is selected to estimate the probability of a pixel belonging to a certain class. The task of the UCN-I is binary classification and we apply a sigmoid activation in the output layer. The UCN-II aims to assign the pixels to the background and five esophageal layer regions with the label from “0” to “4”, so the final channel is set to 5 with softmax activation.

4) LOSS FUNCTION

The loss function for the two UCNs is defined as follows:

$$J = J_{CE} + \lambda_1 J_{Dice} + \lambda_2 \| W \|^2_F$$

where $J_{CE}$ and $J_{Dice}$ represents the cross entropy loss and the Dice loss [29], respectively. The $\| W \|^2_F$ is the Frobenius norm of the weight matrix. $\lambda_1$ and $\lambda_2$ are the user-defined weight parameters to control the trade-off of the three terms. More details are described as follows.

The widely used cross entropy loss function is defined as

$$J_{CE} = -\frac{1}{N} \sum_{i=1}^{N} g_l(x_i) \log p_l(x_i)$$

where $N$ is the total pixel number of the image. $g_l(x_i)$ is the target probability that pixel $x_i$ belongs to class $l$ with one for the true label and zero entries for the others. $p_l(x_i)$ is the corresponding estimated probability of pixel $x_i$ belongs to class $l$. The cross entropy provides a probability similarity between the actual label and the predicted value of the network.

Eq. (2) tells that the pixels of each class contribute equally to the cross entropy loss. However, the case in the current study has severe class imbalance problem, which means the anatomical structures occupy only small regions of the whole scan as shown in Fig. 1. This may result in the network missing or only partially detecting pixels from the corresponding class. The dice loss is intended to alleviate the issue, which
measures the spatial overlap between the predicted area and the ground-truth. In this case, the dice loss is defined as Eq. (3) [29].

\[
J_{\text{dice}} = 1 - \frac{2 \sum_{i=1}^{N} p_i(x_i) g_i(x_i)}{\sum_{i=1}^{N} p_i^2(x_i) + \sum_{i=1}^{N} g_i^2(x_i)}
\]  

(3)

Detailed definition of Frobenius norm can be found in Eq. (4).

\[
\|W\|_F = \sqrt{\sum_i \sum_j w_{ij}^2}
\]  

(4)

This regularization term is used to prevent overfitting, thus obtaining more robust networks. In this study, \(\lambda_1\) is set at 0.3 and \(\lambda_2\) is set at \(1 \times 10^{-4}\).

C. TRAINING STRATEGIES

1) DATA AUGMENTATION

Massive annotated esophageal OCT images are considered to be inaccessible, which may prevent the network from a satisfactory performance. Data augmentation is an effective way to overcome the sparsity of training dataset [28] and improve the network robustness. The data augmentation techniques used in this study include random rotation, horizontal flipping, random shearing, elastic deformations [30], intensity shifts and multiplicative speckle noise. The first three methods are simple geometric transformations belonging to affine transformations [28], [30]. Elastic deformation plays an important role in biomedical segmentation tasks since tissue deformations are the most common changes in medical images [30]. It is implemented by defining a grid of squares on the image and then morphing each square by methods such as randomly moving its corners while stretching the image inside. This procedure intends to improve the network robustness to esophageal tissue deformation. The intensity shift is implemented by performing a pixel-wise addition of a scalar to the image. This technique is particularly effective to improve the network invariance to intensity inhomogeneity of OCT images. Multiplicative speckle noise is specifically designed to deal with OCT speckle noise that is inherent in coherent imaging. The above data augmentation techniques are randomly applied to each B-scan of the training set. An example of a B-scan undergoing data augmentation is shown in Fig. 5.

2) THE D-UCN TRAINING PROCEDURE

The entire D-UCN network is trained on our dataset using the loss function defined in Eq. (1). The two models of the D-UCN network were trained end-to-end individually using the Adam [34] optimizer with Nesterov momentum of 0.9.
An initial learning rate of $1 \times 10^{-4}$ is applied and is decayed by a factor of 10 if the validation loss fails to improve over ten consecutive epochs. Training is performed in batches of 80 randomly chosen samples at each iteration (selected to saturate the GPU memory). After going through the entire training set, an epoch is finished. Finally, the model with the lowest validation loss is employed to measure the segmentation performance of the testing dataset and for further quantitative evaluation.

Each B-scan from our dataset is of size $1024 \times 1024$. Considering the fact that the target tissue area exists in the upper half of the image, we crop each B-scan along depth to the size of $512 \times 1024$, which is enough to cover all the anatomical information. Then the B-scan is split width-wise into 8 non-overlapped slices (sizing $512 \times 128$) as shown in Fig. 6.

It is worth noting that our network is fully-convolutional, which can be applied to images of arbitrary size, so images are used for model testing without slicing.

### III. EXPERIMENTS

#### A. DATA AND EXPERIMENTAL ENVIRONMENT

Esophageal OCT images from six C57BL mice were used to evaluate the proposed segmentation network. During the experiment, 800 B-scans were collected from four C57BL mice to establish a segmentation network, among which 600 B-scans were randomly selected for training, and the remaining 200 B-scans were used for validation. An independent test set consists of 240 B-scans collected from two other mice, ensuring that there is no overlap between the data used for training and testing.

The annotated labels were generated by an experienced grader using ITK-SNAP [35], which were used for network training and algorithm evaluation. The D-UCNs were implemented in Keras using Tensorflow as the backend. Training of the network was performed on an 11 GB Nvidia GeForce RTX 2080Ti GPU using CUDA 9.2 with cuDNN v7.

#### B. ANALYSIS OF ACCURACY AND LOSS CURVES FOR DIFFERENT MODELS

1) ACCURACY AND LOSS CURVES FOR D-UCN

The proposed D-UCN is a two-stage model. The first stage UCN-I is used to discriminate target tissues and background. The loss and accuracy curves are shown in Fig. 7. We trained UCN-I for 80 epochs and the network reaches an accuracy about 0.99. There is little gap between training and validation curves, indicating the network was not overfitted.

The UCN-II is used to classify different tissues. The loss and accuracy curves are shown in Fig. 8. We trained UCN-II for 100 epochs and the network converges with about 0.99 validation accuracy. The training and validation process
has similar loss and accuracy curves, indicating the training and validation data have similar distributions. This is consistent with the fact that esophageal tissues have similar shapes with the background removed.

2) ACCURACY AND LOSS CURVES FOR U-NET (SINGLE STAGE UCN)
The U-Net used in this study has the same architecture as UCN-II. As a result, the U-Net can be regarded as a single stage UCN. The accuracy and loss curves for U-Net are presented in Fig. 9. The U-Net was trained for 100 epochs and the best validation accuracy reaches 0.98. Comparing with D-UCN, the U-Net has more wavy loss and accuracy curves, indicating it is less robust to train, resulting in more difficulties to achieve a satisfactory model.

3) ABLATION STUDY ON POOLING METHODS
The proposed D-UCN utilized max pooling in the first encoding block and average pooling in other layers. In this case, the network is able to capture informative features from original resolution as well as improve the robustness of the architecture. Fig. 10 shows the changes of accuracy curve when we use only average pooling or max pooling in UCN-II.

Comparing Fig. 10 with Fig. 8(a), we can find that in all three cases, the best verification accuracy is almost equal (about 0.99). The difference lies in the shape of the curve, which reflects the different training processes. Fig. 10(a) with all max pooling takes about 15 epochs to reach a 0.99 validation accuracy, however, the accuracy may meet a sudden drop that is caused by noise or other disturbance since max pooling using part of the input information is more sensitive to noise. On the contrary, Fig. 10(b) will not have those drops, but it takes more epochs to achieve the best validation accuracy. The reason is that average pooling is robust since it uses the average value of the input which sacrifices the ability of capturing informative features from original resolution. The proposed architecture combined advantages of these two pooling methods, which realize fast converge and robust training as shown in Fig. 8(a).

4) ABLATION STUDY ON ITEMS OF LOSS FUNCTION
Fig. 11 shows the accuracy curve of UCN-II that uses only cross entropy loss and dice loss. It can be found that fast convergence is achieved in both cases. However, compared with Fig. 8(a), the segmentation accuracy is much lower. One reason for this is that the esophagus tissue is located in a small area compared to the entire image, which makes it difficult for the algorithm to find the correct optimization direction only based on cross entropy or dice coefficients.
5) ABLATION STUDY ON AUGMENTATION METHODS
In data augmentation, this paper employs elastic deformation and some traditional methods (such as rotation, shearing) to simulate the real appearance of the esophagus OCT image. The deformation sometimes results in unnatural images, which makes people concerned about using it in augmentation. In this case, we add an ablation study to compare the performance of not using deformations in training, and the accuracy curve of UCN-II can be found in Fig. 12. Compared with Fig. 8(a), the highest validation accuracy is almost the same, which confirms that the use of deformations has little effect on average validation accuracy. In this study, the elastic deformation is employed as a data augmentation method to make the network have the potential to process images collected in abnormal conditions.

C. COMPARISON EXPERIMENTS
We compared the proposed D-UCN with two widely adopted segmentation methods, including graph theory dynamic program (GTDP) [13] and U-Net [28]. GTDP is a model-based segmentation method that identifies tissue layer boundaries by optimizing gradient-based graphs. Detailed segmentation procedure can be found in our previous work [16]. The U-Net is a popular pixel classification framework. In this experiment, U-Net is constructed using an architecture similar to UCN-II.

The selected image associated with its corresponding segmented results (manual, GTDP, U-Net and our D-UCN) is shown in Fig. 13. The B-scan displayed in Fig. 13(a) presents the layered esophageal tissue structure. However, some factors that are not conducive to segmentation can also be found in the image, such as the plastic sheath clinging to the tissue, the vague tissue boundary between the last two layers and some irrelevant structures. The results of
FIGURE 13. Demonstration of (a) an esophageal OCT image from the C57BL mouse; (b) label map of manual segmentation; (c) optical staining based on manual segmentation; (d) optical staining based on the segmentation result of GTDP; (e) optical staining based on the segmentation result of U-Net and (f) optical staining based on the segmentation result of D-UCN.

manual segmentation are used as the gold standard for this experiment, as shown in Figs. 13(b) and Fig. 13(c), where the roughness of layer boundaries seems inevitable in the manual pixel-by-pixel annotation.

It can be found from Fig. 13(d) that the esophageal layers delineated by GTDP is inconsistent with the manual annotation. Firstly, the SC layer is much thicker, which is caused by the fact that the plastic sheath adheres to the tissue (marked by red arrow in 13(a)), and limitations of GTDP make it in difficult in distinguishing tissue layers in this situation, thus generating a much thicker “SC”. Secondly, the layer thickness of the last two layers is non-uniform. One reason for this is that the boundary between the last two layers is vague, making it difficult to locate proper pixels of the boundary.

In contrast, the U-Net segmentation result in Fig. 13(e) is less affected by these two problems. However, it may perform segmentation out of the target region since some unrelated tissues may also have layered structure (marked by red arrow in 13(a)). Although the tissues are far from the segmentation target in this example, which seems easy to remove using simple image processing algorithms, the position and shape of such problems are unpredictable. As a result, using another network to locate the target region is a more reasonable choice to avoid segmenting outliers. The D-UCN segmentation of the esophageal tissues in Fig. 13(f) is comparable to the manual segmentation. Benefiting from the dual-stage strategy, the proposed method will focus on the target region, and obtains smooth delineation of esophageal layers.

| Methods | GTDP     | U-Net    | D-UCN     |
|---------|----------|----------|-----------|
| Accuracy| 0.9602 ± 0.0097 | 0.9799 ± 0.0053 | 0.9806 ± 0.0078 |
1) EVALUATION METRICS
We use the following metrics to evaluate the proposed D-UCN framework, including the precision, recall, Dice coefficient, the Hausdorff distance (HD) and the average distance (AVD) [36], [37]. The precision, recall and Dice coefficient evaluated the segmentation performance based on the overlap area, which are defined as Eqs. (5) to (7).

\[
\text{Precision} = \frac{S_R \cap S_G}{S_G} \tag{5}
\]

\[
\text{Recall} = \frac{S_R \cap S_G}{S_R} \tag{6}
\]

\[
\text{Dice} = \frac{2 \times S_R \cap S_G}{S_R + S_G} \tag{7}
\]

where \(S_R\) and \(S_G\) represent the binary segmentation and ground truth areas, respectively. The HD and AVD are used to measure the boundary accuracy of the segmentation result. Definition of HD is given by Eq. (8).

\[
\text{HD}(A, B) = \max(h(A, B), h(B, A)) \tag{8}
\]

where \(A\) and \(B\) are two finite point sets, \(h(A, B)\) is called the directed Hausdorff distance and given by

\[
h(A, B) = \max_{a \in A} \min_{b \in B} \|a - b\| \tag{9}
\]

The AVD is less sensitive to outliers than the HD, which is defined as

\[
\text{AVD}(A, B) = \max(d_d(A, B), d_d(B, A)) \tag{10}
\]

where \(d_d(A, B)\) is called the directed Average Hausdorff distance given by

\[
d_d(A, B) = \frac{1}{N} \sum_{a \in A} \min_{b \in B} \|a - b\| \tag{11}
\]

2) QUANTITATIVE ASSESSMENT OF THE D-UCN
The evaluation is implemented with the testing dataset consisting of 240 B-scans. The overall segmentation accuracy of different methods for the testing dataset is shown in Table 1. The proposed D-UCN achieved the highest segmentation accuracy. However, for the reason that the background is much larger than target tissues, different labeling of target tissues will not bring significant changes in the overall accuracy. Therefore, D-UCN did not present obvious advantages over the other methods.

More detailed analysis of different methods for the four target tissue layers (SC, EP, LP&MM, SM) are listed in Tables 2 to 5 with the best performance bolded. From Table 2, we find that the GTDP with a significantly lower precision value and Dice coefficient for the “SC” layer, which means the GTDP is less effective in segmenting this part. The reason for this has been explained in the earlier section. Compared with GTDP, the U-Net results show better agreement with manual segmentation, with higher precision, recall and Dice values for all tissues. However, its HD is abnormally large, indicating the segmentation result has some outliers as Fig. 13(e) suggests. Overall, the D-UCN performed better for all the tissues in comparison with the GTDP-based results. Moreover, the D-UCN has significant improvement in HD compared with the U-Net-based result, indicating the two-stage strategy is a solution to avoid the problem caused by irrelevant structures. In addition, the segmentation accuracy is also improved as demonstrated by the tables.

### IV. DISCUSSIONS
Automatic tissue layer segmentation is the crucial step for esophageal morphology analysis of endoscopic OCT images. However, our previous research found that it is challenging to use traditional methods to accomplish this task without manual intervention. A major reason is that the collected esophageal OCT images contain a large proportion of adjacent tissues neighboring to the layers of interest together with the plastic sheath adjoining the esophagus that has a negative effect on precise layer segmentation. Our D-UCN framework combines a priori knowledge of esophageal anatomy to solve this problem effectively, which first extracts target tissue area in the first stage and then implements esophageal layer segmentation in the second stage. The D-UCN is trained and validated on the collected esophageal OCT images and testing is performed on an independent test set. Experimental results show that the proposed method can achieve performance...
TABLE 4. Evaluation of different segmentation methods for the LP&MM layer.

| Methods | Precision | Recall  | Dice    | HD (μm)  | AVD (μm)  |
|---------|-----------|---------|---------|----------|-----------|
| GTDP    | 0.7795 ± 0.0851 | 0.7111 ± 0.1025 | 0.7410 ± 0.0839 | 4.8339 ± 2.7676 | 0.8660 ± 0.4573 |
| U-Net   | 0.8641 ± 0.0576 | 0.8340 ± 0.1226 | 0.8447 ± 0.1018 | 12.6476 ± 16.3150 | 1.2714 ± 1.3176 |
| D-UCN   | 0.8764 ± 0.0490 | 0.8480 ± 0.0258 | 0.8607 ± 0.0199 | 3.2703 ± 3.5382 | 0.5019 ± 0.1346 |

TABLE 5. Evaluation of different segmentation methods for the SM layer.

| Methods | Precision | Recall  | Dice    | HD (μm)  | AVD (μm)  |
|---------|-----------|---------|---------|----------|-----------|
| GTDP    | 0.7913 ± 0.0500 | 0.7906 ± 0.0662 | 0.7903 ± 0.0549 | 4.5665 ± 3.4754 | 0.7834 ± 0.4931 |
| U-Net   | 0.8709 ± 0.0683 | 0.7992 ± 0.1411 | 0.8278 ± 0.1298 | 24.5274 ± 16.0825 | 1.8135 ± 1.7129 |
| D-UCN   | 0.8201 ± 0.0542 | 0.8908 ± 0.0325 | 0.8523 ± 0.0267 | 3.2939 ± 3.8266 | 0.4856 ± 0.1186 |

FIGURE 14. Demonstration of (a) an esophageal OCT image with discontinuous tissues and (b) the corresponding segmentation result of D-UCN.

Comparability to manual segmentation. In addition, we designed experiments to use more convincing quantitative evaluation methods to evaluate the improvement of the added target region extraction step.

In terms of network architecture design, we utilized techniques such as skip connections, residual learning, batch normalization, average pooling to improve the network performance. The advantages of the mentioned techniques have been reported in the literature [28], [38], [39]. However, assembling those approaches to get a best architecture is challenging. In this work, we first use different combinations of these methods to build the architecture, and then calculate the training loss and validation accuracy of different architectures on the same training and validation dataset. Finally, we chose the architecture with the highest validation accuracy to achieve the proposed UCN. To further improve the UCN architecture, hyperparameter optimization [40] or ensemble learning [41] may be a choice. However, the improvement may not be that significant considering the computation cost.

Esophageal tissues are continuous in healthy condition. However, an abrupt gap sometimes occurred on tissues, which is caused by imaging equipment or diseases. An example of esophageal OCT images with gaps on tissues generated by unfavorable imaging condition is shown in Fig. 14(a). It is noticed from Fig. 14(b) that the D-UCN is able to segment part of the missing tissues based on the existing structure, while for tissues that cannot be inferred, D-UCN will just segment the continuous tissues and label the missing part as background.

In the current study, the network was trained with esophageal OCT images collected from mice. For endoscopic OCT images containing esophageal diseases, or making the move from analyzing dataset of lab animals to human diagnosis, we should retrain the D-UCN with extended data.

V. CONCLUSION

This study proposes a new D-UCN framework that can adaptively segment four esophageal tissue layers from different imaging conditions. The D-UCN consists of two cascaded UCNs specifically designed for automatic segmentation of esophageal layers on OCT images. The network architecture combines several recently developed techniques, such as skip connections, residual learning, batch normalization and average pooling to improve the network performance. The first UCN aims to locate the esophageal anatomy regions. Its output is used as a priori knowledge to assist the second UCN for accurate layer segmentation. Training of the network was implemented by optimizing a synthesized loss function combining cross entropy loss and dice loss. Experiments on mouse esophageal segmentation proved that the dual-stage network with prior target region knowledge is more effective than GTDP and U-Net. Leveraging on the inherent advantages of a fully convolutional network, the application of D-UCN to other esophageal diseases is not difficult. It is also
appealing to extend the D-UCN to study OCT images from other organs such as the airway and retina.

REFERENCES

[1] P. S. Elias and D. O. Castell, “The role of acid suppression in Barrett's esophagus,” *Amer J. Med.*, vol. 130, no. 5, pp. 525–529, May 2017.

[2] P. N. Thota, G. Zuccaro, Jr., J. J. Vargo, H. L. D. Cornelj, J. A. Dumot, and M. Yu, “A randomized prospective trial comparing unseeded esophagoscopy via transnasal and transoral routes using a 4-mm video endoscope with conventional endoscopy with sedation,” *Endoscopy*, vol. 37, no. 6, pp. 559–565, Jun. 2005.

[3] J. Melson, V. Desai, M. Greenspan, S. Yau, M. Abdalla, R. Dhaneokula, S. Mobarhan, D. Shapiro, J. Losuro, and S. Jakate, “Negative surveillance endoscopy occurs frequently in patients with short-segment non-dysplastic Barrett’s esophagus,” *Diseases Esophagus*, vol. 28, no. 7, pp. 660–665, Oct. 2015.

[4] D. Huang, E. A. Swanson, C. P. Lin, J. S. Schuman, W. G. Stinson, W. Chang, M. R. Hee, T. Flotte, K. Gregory, C. A. Puliafito, and J. G. Fuhimoto, “Optical coherence tomography,” *Science*, vol. 254, no. 5035, pp. 1178–1181, 1991.

[5] G. J. Tearney, M. E. Brezinski, B. E. Bouma, S. A. Boppart, C. Pitris, J. F. Southern, and G. Fuhimoto, “**vivo** endoscopic optical coherence tomography,” *Science*, vol. 276, no. 5321, pp. 2037–2039, Jun. 1997.

[6] J. Xi, A. Zhang, Z. Liu, W. Liang, L. Y. Lin, S. Yu, and X. D. Li, “Diffractive catheter for ultra-high-resolution spectral-domain volumetric OCT imaging,” *Opt. Lett.*, vol. 39, no. 7, pp. 2016–2019, Apr. 2014.

[7] J. A. Dumot, and M. Xu, “A randomized prospective trial comparing no. 9, pp. 4481–4495, 2018.

[8] S. Yuan, and Q. Chen, “Automated 3-D retinal layer segmentation from SD-OCT images with neurosensory retinal detachment,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 152388–152398, 2019.

[9] J. L. Zhang, W. Yuan, W. X. Liang, S. Y. Yu, Y. M. Liang, Z. Y. Xu, J. A. Dumot, and M. Xu, “**vivo** endoscopic optical coherence tomography,” *Science*, vol. 254, no. 5321, pp. 1429–1433, Dec. 2006.

[10] Z. Liu, J. Xi, M. Tse, A. C. Myers, X. Li, P. J. Pasricha, and S. Yu, “426 allergic inflammation-induced structural and functional changes in esophageal epithelium in a guinea pig model of eosinophilic esophagitis,” *Gastroenterology*, vol. 120, no. 1, pp. 7–12, Jan. 2001.

[11] S. H. Yun, G. J. Tearney, B. J. Vakoc, M. Shishkov, W. Y. Oh, A. E. Desjardins, M. J. Suter, R. C. Chan, J. A. Evans, I.-K. Jang, N. S. Nishioka, J. F. de Boer, and B. E. Bouma, “Comprehensive volumetric optical microscopy in vivo,” *Nature Med.*, vol. 12, no. 12, pp. 1429–1433, Dec. 2006.

[12] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and P. J. Pasricha, “**vivo** endoscopic optical coherence tomography and laser cautery marking: A pilot clinical study,” *Gastrointestinal Endoscopy*, vol. 79, no. 6, pp. 886–896, Jun. 2014.

[13] G. J. Ughi, M. J. Gora, A. F. Swager, A. Soomro, C. Grant, A. Tiernan, M. Rosenberg, J. S. Sauk, N. S. Nishioka, and G. J. Tearney, “Automated segmentation and characterization of esophageal wall in vivo by tethered capsule optical coherence tomography endomicroscopy,” *Biomed. Opt. Exp.*, vol. 7, no. 2, pp. 409–419, 2016.

[14] J. L. Zhang, W. Yuan, W. X. Liang, S. Y. Yu, Y. M. Liang, Z. Y. Xu, Y. X. Wei, and X. D. Li, “Automatic and robust segmentation of endoscopic OCT images and optical staining,” *Biomed. Opt. Exp.*, vol. 8, no. 5, pp. 2697–2708, 2017.

[15] S. J. Chiu, X. T. Li, P. Nicholas, C. A. Toth, J. A. Izatt, and S. Farsi, “Automatic segmentation of seven retinal layers in SD-OCT images congruent with expert manual segmentation,” *Opt. Exp.*, vol. 18, no. 18, pp. 19413–19428, 2010.

[16] B. I. Dodo, Y. Li, D. Kaba, and X. Liu, “Retinal layer segmentation in optical coherence tomography images,” *IEEE Access*, vol. 7, pp. 152388–152398, 2019.

[17] L. Bekalo, S. Niu, X. Pe, P. Li, P. Okwobi, C. Yu, W. Fan, S. Yuan, and Q. Chen, “Automated 3-D retinal layer segmentation from SD-OCT images with neurosensory retinal detachment,” *IEEE Access*, vol. 7, pp. 14894–14907, 2019.

[18] M. Gan, C. Wang, T. Yang, N. Yang, M. Zhang, W. Yuan, X. D. Li, and L. R. Wang, “Robust layer segmentation of esophageal OCT images based on graph search using edge-enhanced weights,” *Biomed. Opt. Exp.*, vol. 9, no. 9, pp. 4481–4495, 2018.
[36] A. A. Taha and A. Hanbury, “Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool,” BMC Med. Imag., vol. 15, no. 1, p. 29, Dec. 2015.

[37] H. Zhao, B. He, Z. Ding, K. Tao, T. Lai, H. Kuang, R. Liu, X. Zhang, Y. Zheng, J. Zheng, and T. Liu, “Automatic lumen segmentation in intravascular optical coherence tomography using morphological features,” IEEE Access, vol. 7, pp. 88859–88869, 2019.

[38] H. Manukian, F. L. Traversa, and M. Di Ventra, “Accelerating deep learning with memcomputing,” Neural Netw., vol. 110, pp. 1–7, Feb. 2019.

[39] S. Mehrkanoon, “Deep neural-kernel blocks,” Neural Netw., vol. 116, pp. 46–55, Aug. 2019.

[40] L. A. Thiede and U. Parlitz, “Gradient based hyperparameter optimization in echo state networks,” Neural Netw., vol. 115, pp. 23–29, Jul. 2019.

[41] X.-L. Zhang, “Multilayer bootstrap networks,” Neural Netw., vol. 103, pp. 29–43, Jul. 2018.

MENG GAN received the Ph.D. degree from the University of Science and Technology, Hefei, China, in 2017. From 2017 to 2020, she was a Postdoctoral Researcher with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. The principal aim of her postdoctoral project is to establish an automatic image segmentation system for analyzing images collected from a high-resolution endoscopic OCT system. Since 2020, she has been a Research Assistant with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. Her research interests include optical coherence tomography imaging, medical image processing, and deep learning.

CONG WANG was born in Yanji, Jilin, China, in 1991. He received the B.S. degree and the Ph.D. degree in mechanical engineering from the University of Science and Technology, Hefei, China, in 2012 and 2017, respectively. From 2017 to 2020, he was a Postdoctoral Researcher with the Department of Medical Diagnostic, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. Since 2020, he has been a Research Assistant with the Department of Medical Optics, Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. His research interests include wavelet transform, medical image processing, and application of deep learning techniques in computer vision.

* * *

MENG GAN received the Ph.D. degree from the University of Science and Technology, Hefei, China, in 2017. From 2017 to 2020, she was a Postdoctoral Researcher with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. The principal aim of her postdoctoral project is to establish an automatic image segmentation system for analyzing images collected from a high-resolution endoscopic OCT system. Since 2020, she has been a Research Assistant with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences. Her research interests include optical coherence tomography imaging, medical image processing, and deep learning.