






Received August 14, 2020, accepted August 20, 2020, date of publication August 26, 2020, date of current version September 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019518

Eso-Net: A Novel 2.5D Segmentation Network With the Multi-Structure Response Filter for the Cancerous Esophagus

DONGHAO ZHOU¹, GUOHENG HUANG¹, (Member, IEEE), JIAJIAN LI², SIYU ZHU³, ZHUOWEI WANG¹, BINGO WING-KUEN LING¹, (Senior Member, IEEE), CHI-MAN PUN⁴, (Senior Member, IEEE), LIANGLUN CHENG², (Senior Member, IEEE), XIUYU CAI⁵, AND JIAN ZHOU³

¹School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China

²School of Computers, Guangdong University of Technology, Guangzhou 510006, China

³Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

⁴Department of Computer and Information Science, University of Macau, Zhuhai, Macau

⁵Department of Medical Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China

Corresponding authors: Guoheng Huang (kevinwong@gdut.edu.cn), Xiuyu Cai (caixy@sysucc.org.cn), and Jian Zhou (zhoujian@sysucc.org.cn)


This work was supported in part by the National Natural Science Foundation of China under Grant 61702111; in part by the National Key Research and Development Program of China under Grant 2016YFC0800506 and Grant 2017YFB1201203; in part by the National Nature Science Foundation of China-Guangdong Joint Fund under Grant 83-Y40G33-9001-18/20; in part by the Guangdong Esophageal Cancer Institute Science and Technology Program under Grant Q-201602 and Grant Q-201607; in part by the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2016B030301008; in part by the National Natural Science Foundation of Guangdong Joint Fund under Grant U1801263 and Grant U1701262; in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2018B010109007, Grant 2019B010109001, and Grant 2019B010153002; and in part by the Blue Fire Plan (Huizhou) Industry-University-Research Joint Innovation Fund 2017 Project of the Ministry of Education under Grant CXZJHZ201730.

ABSTRACT Automatic segmentation of the cancerous esophagus in computed tomography (CT) images is a computer-assisted method that can improve the efficiency of the diagnosis and treatment. Due to the diversity of the cancer stage and location, the anatomical structure of the cancerous esophagus is various. Moreover, the low contrast against surrounding tissues leads to a blurry boundary of the cancerous esophagus. Therefore, existing segmentation networks cannot achieve satisfactory results in automatic segmentation of the cancerous esophagus. In this article, we propose a novel 2.5D segmentation network named Eso-Net for the cancerous esophagus based on an encoder-decoder architecture. A 3D enhancement filter called Multi-Structure Response Filter (MSRF) is designed to extract 3D structural information as prior knowledge. Furthermore, dilated convolutions and residual connections are employed in the convolutional blocks of Eso-Net for multi-scale feature learning. With 3D structural priors, Prior Attention Modules (PAM) are incorporated into the network to facilitate the transmission of relevant spatial information. The experiments are conducted on the dataset from 30 esophageal cancer patients, and we report an 84.839% dice similarity coefficient, an 85.955% precision, an 83.752% sensitivity, and a 2.583mm Hausdorff distance. The experimental results demonstrate that the proposed method outperforms other existing segmentation networks in this task and can effectively assist doctors in the diagnosis and treatment of esophageal cancer.

INDEX TERMS Esophageal cancer, medical image segmentation, deep learning, attention mechanism, enhancement filter.

I. INTRODUCTION

Esophageal cancer is a common cancer with a high mortality [1], and becomes a major public health problem worldwide since its incidence has been increasing in

The associate editor coordinating the review of this manuscript and approving it for publication was K. C. Santosh .

recent years [2]. Therefore, the diagnosis and treatment of esophageal cancer is particularly critical. Medical imaging is a technique for the diagnosis and clinical analysis of diseases since it can reveal internal structures of a body. Among many imaging methods, Computed Tomography (CT) is widely used for the diagnosis of esophageal cancer, since it can create visual representations of body cross-sections to assist doctors

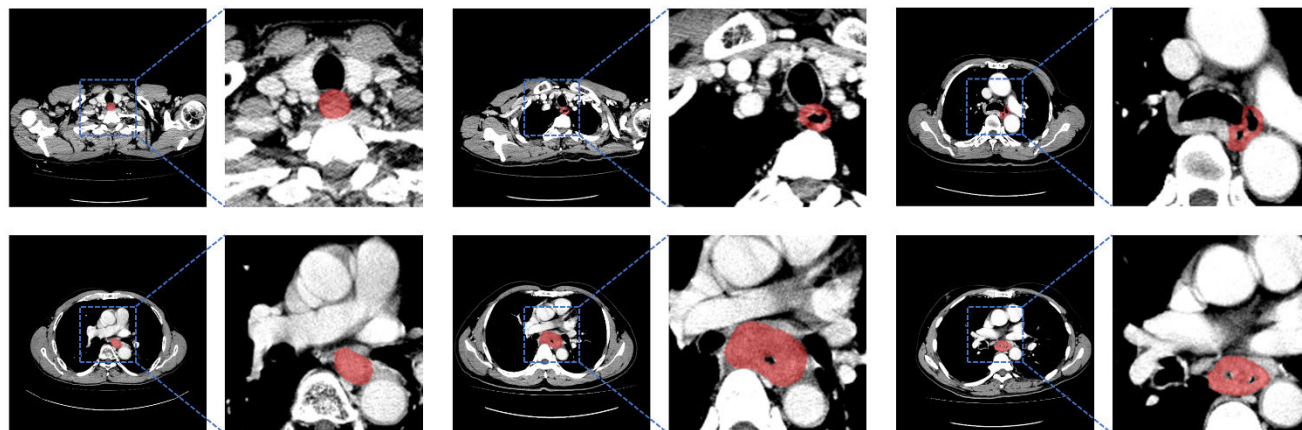


FIGURE 1. Some typical CT slices. The esophagus regions are indicated in red.

to observe and evaluate the esophageal lesions [3]. However, manual segmentation of the cancerous esophagus is tedious and time-consuming due to the large number of CT slices. Limited by professional abilities of doctors, some segmentation results may be subjective or even inaccurate, which has serious consequences. Therefore, automatic methods are developed for the accurate and effective segmentation of the cancerous esophagus, which can assist doctors to diagnose esophageal cancer. However, automatic segmentation of the cancerous esophagus in CT images is still a difficult task. As shown in Fig. 1, the size and shape of the cancerous esophagus is complex and variable due to the diversity of the cancer stage and location. Moreover, the boundary of the esophagus is irregular and blurry in CT images since the esophagus is a non-rigid structure with low contrast against surrounding tissues. Furthermore, air bubbles randomly appear in the esophagus, which further increases the difficulty of segmentation.

Using traditional image processing, many methods have been proposed for this task. Rousson *et al.* described a probabilistic shortest path approach combined with spatial prior knowledge [4]. This method models spatial dependency between the esophagus and the aorta and left atrium to obtain the esophagus outer boundary. However, this method requires the two extreme points of the esophagus centerline and a segmentation of the aorta and left atrium as input. Feulner *et al.* proposed a multi-step method to extract the esophagus from CT images [5]. A classifier for discriminating appearances is combined with an explicit model of the respiratory and esophageal air distribution, followed by a Markov chain model to estimate the approximate esophagus shape. Then the approximate surface performs non-rigid deformation to obtain a better fitting boundary. In [6], Damien *et al.* introduced a skeleton-shaped model to guide the segmentation. This method performs a 3D segmentation with the prior knowledge of the skeleton. Then over-segmented slices are detected and a 2D propagation by graph cut is used to improve segmentation results. Yang *et al.* proposed an online atlas

selection approach for multi-atlas segmentation [7]. Based on local anatomical similarity, the atlases of the esophagus are ranked and the optimal atlases are selected. The final segmentation is obtained by fusing the deformed contours of the optimal atlases. Since most methods using traditional image processing are proposed for specific scenarios with pre-defined hypotheses, the generalization capabilities of the algorithms are restricted. Moreover, the feature extractors of the esophagus are artificially designed with complex parameters, which reduces the robustness and requires tedious parameter tuning.

Recently, deep learning has been widely used in image segmentation. Shelhamer *et al.* proposed a pixels-to-pixels model called Fully Convolutional Networks (FCN) [8], based on an encoder-decoder architecture without any fully connected layers. Fechter *et al.* proposed a random walker approach driven by a 3D FCN [9], which is believed as the first work to apply deep learning in the segmentation of the esophagus. This method is not end-to-end since the 3D FCN is just used to generate a rough probability map of the esophagus. Trullo *et al.* utilized two improved FCN to perform and refine the segmentation of the esophagus [10]. However, a multi-organ segmentation is required to locate the esophagus, which complicates the model training and data labeling. Furthermore, some novel segmentation networks are developed based on FCN. U-Net combines a U-shaped network with skip connections to fuse features and recover lost spatial information [11]. To achieve less memory requirement and inference time overhead, SegNet uses max-pooling indices to perform non-linear up-sampling [12]. LinkNet connects each encoder input to the corresponding decoder output for recovering lost spatial information [13]. Due to the compact but efficient network topology, U-Net becomes the focus of biomedical image segmentation. Çiçek *et al.* employed 3D convolutions to construct 3D U-Net with a large number of trainable parameters that can easily lead to the model overfitting without a large amount of data [14]. Chen *et al.* proposed U-Net plus to segment the cancerous esophagus

on a single CT slice [15]. However, this method requires a manually placed point to start segment the entire esophagus.

Even using deep learning, automatic segmentation of the cancerous esophagus remains a challenging problem. The networks mentioned above also failed to achieve satisfactory results, which can be described in three aspects. First, limited by GPU memory, 3D segmentation can just process a small patch of the CT scan, which is not conducive to learning the complete esophagus structure. Whereas 2D segmentation cannot utilize 3D structural information. Second, various tumor sizes lead to various scales of the cancerous esophagus. It is difficult to achieve multi-scale feature learning in this task simply by using down-sampling to enable convolutions to extract multi-scale features. Third, the loss of spatial information in the encoding phase is one of the major factors that limit segmentation accuracy. Skip connections of U-Net is a solution by simply concatenating the feature maps from decoder layers and encoder layers. However, it is not well enough to recover and fuse lost spatial information of the cancerous esophagus that has an irregular and blurry boundary. To solve the above problems, we propose a novel segmentation network named Eso-Net that consists of different convolutional blocks. Eso-Net performs channel-wise 2.5D segmentation with 3D structural priors extracted by a 3D enhancement filter, which is a trade-off between 2D and 3D segmentation. In medical image tasks, the term “2.5D” means that the network utilizes only 2D convolutions to extract features from multiple images in different planes or different views. In the convolutional blocks, dilated convolutions and residual connections are applied to extract and combine features in different receptive fields. Furthermore, the encoder blocks and decoder blocks are connected and an attention mechanism is introduced to make skip connections more effective.

The contributions of our work are summarized as follows:

- 1) We design a robust 3D enhancement filter called Multi-Structure Response Filter (MSRF) to extract 3D structural information that can instruct the network to distinguish the cancerous esophagus. For the balance between the model size and segmentation accuracy, channel-wise 2.5D segmentation is performed with the assistance of 3D structural priors.
- 2) We propose Eso-Net for automatic and efficient segmentation of the cancerous esophagus. Dilated convolutions and residual connections are employed to construct the convolutional blocks of Eso-Net, which is conducive to multi-scale feature extraction and fusion without requiring complex architecture.
- 3) To emphasize the esophagus and suppress irrelevant tissues and organs, we design the Prior Attention Module (PAM) embedded in the connection paths between the encoder blocks and decoder blocks. Moreover, 3D structural priors are utilized to supervise the transmission of spatial information in skip connections.
- 4) According to the experiments, the proposed method obtains the best performance in this task with the

highest dice similarity coefficient (84.839%), precision (85.955%), sensitivity (83.752%), and the lowest Hausdorff distance (2.583mm). The segmentation results achieve the accuracy needed in practical clinical applications.

II. RELATED WORKS

A. ENHANCEMENT FILTERS FOR MEDICAL IMAGES

Enhancement filters, which are used to detect the tissues with a given structure, have been a research hotspot in the medical image field. Based on the Hessian matrix, Frangi *et al.* designed a vessel enhancement filter called Frangi filter and achieved a good effect [16]. Sato *et al.* proposed a series of 3D local structure filters based on a multi-dimensional opacity function [17]. To avoid false positives results, Li *et al.* developed three selective enhancement filters for the dot, line, and plane [18]. Furthermore, some variants of the Frangi filter are employed in the detection of different tissues. Jimenez-Carretero *et al.* proposed a lung vessel filter and applied a penalty function to decrease the filter response in airways [19]. Shahzad *et al.* used the Frangi filter to detect the centerline of subcutaneous veins [20]. Shahid *et al.* proposed a robust method for retinal vessel segmentation combined with the Frangi filter [21]. In deep learning, image enhancement can be considered as a preprocessing approach. Jiang *et al.* used the Frangi filter to enhance the patches of the lung images followed by a four-channel convolution neural network to detect nodules of four levels [22]. Blaiech *et al.* employed the multi-scale technique with the Frangi filter to improve the segmentation performance in the presence of noise [23].

These methods utilized to directly enhance the target tissues are not applicable to the cancerous esophagus since it has various anatomical structures. Anatomical structure is an important property to distinguish different tissues and organs. Inspired by [16], we design a filter to enhance the surrounding tissues and organs of the esophagus, and the enhanced images are used to improve the segmentation accuracy.

B. MULTI-SCALE FEATURE EXTRACTION AND FUSION FOR IMAGE SEGMENTATION

The features extracted from different layers contain different information. The low-level features contain more detail information, whereas the high-level features contain more semantic information. The reason is that down-sampling operations in the network gradually reduces the size of feature maps and convolution kernels can scan at larger scales. In image segmentation, some attempts at multi-scale feature learning have achieved effective results. Yu *et al.* firstly employed dilated convolutions to aggregate multi-scale features without losing resolution [24]. PSPNet uses the pyramid pooling module to extract and concatenate the features of different sizes [25]. Fu *et al.* constructed multi-scale input layers to combine information from multiple scales for accurate segmentation of retinal vessels [26]. Wang *et al.* proposed a hybrid dilated convolution (HDC) framework to alleviate

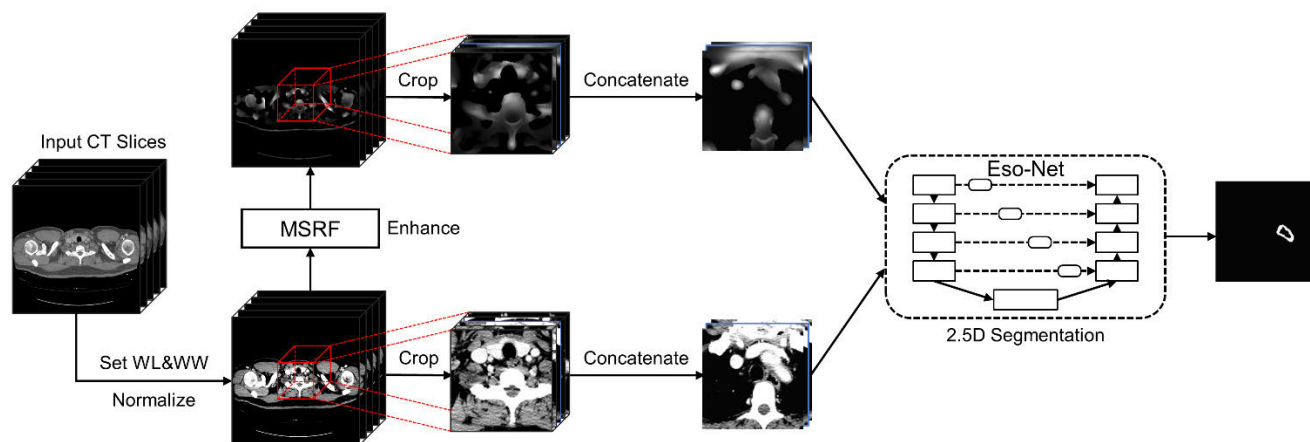


FIGURE 2. The flow chart of the proposed method. The red cubes represent the cropped regions. The images with a blue border are the images to be segmented.

the “gridding issue” caused by dilated convolutions [27]. Chen *et al.* proposed Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale features with dilated convolutions at multiple sampling rates [28]. Zhao *et al.* developed the image cascade network for real-time semantic segmentation by efficiently utilizing information from images with different resolutions [29]. U-Net++ integrates U-Nets of varying depths to extract and aggregate features of varying scales with skip connections [30].

The various scale of the cancerous esophagus requires an effective method for capturing contextual information. Limited by the small size of the cancerous esophagus dataset, the model with a large number of trainable parameters can easily cause overfitting. Excessively employing various network components for feature extraction and fusion is not applicable to this task. Therefore, we subtly combine dilated convolutions and residual connections in the convolutional blocks of our Eso-Net to achieve multi-scale feature learning.

C. ATTENTION MECHANISMS IN MEDICAL IMAGE TASKS

Attention mechanisms were initially proposed for machine translation, and are widely used in the computer vision field [31]–[33]. Attention mechanisms can make the model more focused on key information relevant to the foreground, and has found applications in medical image tasks recently. Attention U-Net combines attention gates with U-Net to emphasize useful features with minimal computational overhead [34]. Roy *et al.* proposed three attention modules modified from the squeeze-and-excitation module and embedded them in different segmentation network to perform multi-organ segmentation [35]. Wang *et al.* proposed global aggregation blocks with a spatial attention mechanism to extract global information of feature maps [36].

All the above are self-attention mechanisms without supervision by prior knowledge. In medical image segmentation, 3D structural information can be used as prior knowledge. Therefore, we utilize 3D structural priors to explicitly guide

the execution of a spatial attention mechanism in Prior Attention Modules (PAM), which contributes to the recovery and fusion of key spatial information. Our PAMs can efficiently model the importance of different regions on a CT image instead of self-learning from feature maps, which facilitates the transmission of features relevant to the esophagus.

III. PROPOSED METHOD

A. OVERVIEW

The proposed method comprises two stages, which are illustrated in Fig. 2. In the first stage, pre-processing is performed on the input CT slices. Firstly, the window level (WL) and window width (WW) of the CT slices are set to 40 and 200 respectively to increase the contrast between the esophagus and other tissues and organs. Intensity values between -160 and 240 are linearly normalized into $[0,1]$ to accelerate the model convergence during training. Intensities less than -160 are set to 0 and those greater than 240 are set to 1. Then, the Multi-Structure Response Filter (MSRF) is used to enhance tissues and organs with specific geometrical structures at multiple scales on CT images, which can provide the segmentation network with additional prior knowledge. Finally, the 144×144 pixels regions in the centers of the original images and the enhanced images are cropped to obtain the Regions of Interest (ROI) that are large enough to contain the entire esophagus. Furthermore, the cropping operation also decreases the interference from irrelevant tissues and organs and speeds up model training and inferencing. In the second stage, channel-wise 2.5D segmentation is performed by Eso-Net. The CT image to be segmented is performed channel-wise concatenation with two adjacent images as one of the network inputs, which is conducive to efficiently utilize z-axis information without a significant increase in the number of parameters. If the image to be segmented is the first or last image of the CT scan, unavailable adjacent images are replaced by the image to be segmented itself. Moreover, another input is the corresponding enhanced images that are

TABLE 1. The Relations between Local Geometrical Structures and the Eigenvalues of the Hessian Matrix.

Structure	λ_1	λ_2	λ_3
plate-like (brighter)	L	L	-H
plate-like (darker)	L	L	+H
tube-like (brighter)	L	-H	-H
tube-like (darker)	L	+H	+H
blob-like (brighter)	-H	-H	-H
blob-like (darker)	+H	+H	+H
no evident structure (background)	L	L	L

H = high and L = low magnitude. +/- indicates the sign of the eigenvalue. "Brighter" and "darker" represent that the structure is brighter and darker than the background respectively, which depends on the intensity of the background.

concatenated in the same way. Finally, the network outputs the segmentation map for the CT image in the middle channel.

In the following, the proposed method is described in detail. In Section III.B., we illustrate the Multi-Structure Response Filter (MSRF). In Section III.C., we introduce Eso-Net that comprises different convolutional blocks and the Prior Attention Module (PAM).

B. MULTI-STRUCTURE RESPONSE FILTER

In medical image segmentation, image enhancement contributes to improve segmentation performance. A common approach to performing image enhancement on CT images is utilizing 3D structural information. Since the size and shape of the esophageal tumor are various, the anatomical structure of the cancerous esophagus is irregular. Therefore, direct enhancement of the cancerous esophagus cannot achieve a satisfactory result. However, other tissues and organs surrounding the esophagus are regular geometrical structures. Inspired by this insight, we propose the Multi-Structure Response Filter (MSRF) based on the Hessian matrix. In contrast to traditional approaches, MSRF is designed to enhance other tissues and organs with specific geometrical structures instead of the esophagus. Thus, the enhanced regions in a CT image should be predicted as the background by the segmentation network. The enhanced images can instruct the segmentation network to distinguish the cancerous esophagus from similar tissues and organs in CT images.

Since the Hessian matrix are related to local geometrical structures [16], we can detect specific structures by using the eigenvalues of the Hessian matrix. Firstly, we combine consecutive CT images into a 3D volume data. The Hessian matrix of each voxel in the 3D volume data comprises second order derivatives in different directions. Let $I(\mathbf{x})$ denotes the intensity of the 3D volume data at coordinate $\mathbf{x} = [x_1, x_2, x_3]^T$. For analyzing structures of multiple scales, differentiation is performed on a Gaussian scale space. Therefore, using the linear scale-space theory [37], the elements in the Hessian matrix of \mathbf{x} at scale σ is defined as:

$$H_{ij}(\mathbf{x}, \sigma) = \sigma^\gamma I(\mathbf{x}) * \frac{\partial^2}{\partial x_i \partial x_j} G(\mathbf{x}, \sigma) \tag{1}$$

where $i, j = 1, 2, 3$ indicate the positions of the elements and $*$ denotes the convolution operation. The parameter γ is introduced to rescale the response of differential operation at

multiple scales [38] and is set to 2. The Gaussian function $G(\mathbf{x}, \sigma)$ is defined as:

$$G(\mathbf{x}, \sigma) = \frac{1}{(2\pi\sigma^2)^{3/2}} \exp\left(-\frac{\mathbf{x}^T \mathbf{x}}{2\sigma^2}\right) \tag{2}$$

The second derivative of a Gaussian kernel at scale σ can be considered as a probe kernel that can capture the difference between the regions inside and outside the range $(-\sigma, \sigma)$ in the direction of the derivative. Moreover, the half-width of the kernel is set to the integer closest to 3σ . when performing the image convolution. Each voxel in the 3D volume data corresponds to a 3×3 Hessian matrix.

The eigenvalue λ_1, λ_2 , and λ_3 are obtained by using eigenvalue decomposition of the Hessian matrix and are sorted in order of absolute values, which means $|\lambda_1| \leq |\lambda_2| \leq |\lambda_3|$. Their magnitudes represent the curvature in the direction pointed by the corresponding eigenvectors since the Hessian matrix describes second order structural characteristics. For a local structure at a specific scale, eigenvalue decomposition extract three representative and orthogonal eigenvectors. Therefore, local geometrical structures can be interpreted by analyzing the signs and magnitudes of λ_1, λ_2 and λ_3 , which is shown in Table 1. For example, in a tube-like structure, the eigenvector corresponding to λ_1 points in the axial direction that is the direction of minimal curvature ($|\lambda_1| \approx 0$) and the other two eigenvectors point in the radial direction that is the direction of larger curvature ($|\lambda_2| \approx |\lambda_3| \gg 0$). Whereas the curvature is large in all directions in a blob-like structure ($|\lambda_1|, |\lambda_2|, |\lambda_3| \gg 0$).

For our purpose, not all evident structures need to be enhanced. MSRF should enhance surrounding tissues and organs with the plate-like, tube-like, and blob-like structure. Moreover, it should avoid enhancing the esophagus and the background. Hence, the enhancement function $E_\sigma(\mathbf{x})$ at scale σ is defined as:

$$E_\sigma(\mathbf{x}) = \begin{cases} 0, & \text{if } \lambda_2 > 0 \text{ or } \lambda_3 > 0, \\ \exp\left(-\frac{R_t^2}{2\alpha^2}\right) \left(1 - \exp\left(-\frac{R_s^2}{2\delta^2}\right)\right), & \\ \exp\left(-\frac{R_p^2}{2\beta^2}\right) \exp\left(-\frac{R_b^2}{2\gamma^2}\right) \left(1 - \exp\left(-\frac{R_s^2}{2\delta^2}\right)\right), & \text{if } \lambda_2, \lambda_3 < 0 \text{ and } \sigma \in S_\sigma, \\ \exp\left(-\frac{R_p^2}{2\beta^2}\right) \exp\left(-\frac{R_b^2}{2\gamma^2}\right) \left(1 - \exp\left(-\frac{R_s^2}{2\delta^2}\right)\right), & \text{if } \lambda_2, \lambda_3 < 0 \text{ and } \sigma \in S_t. \end{cases} \tag{3}$$

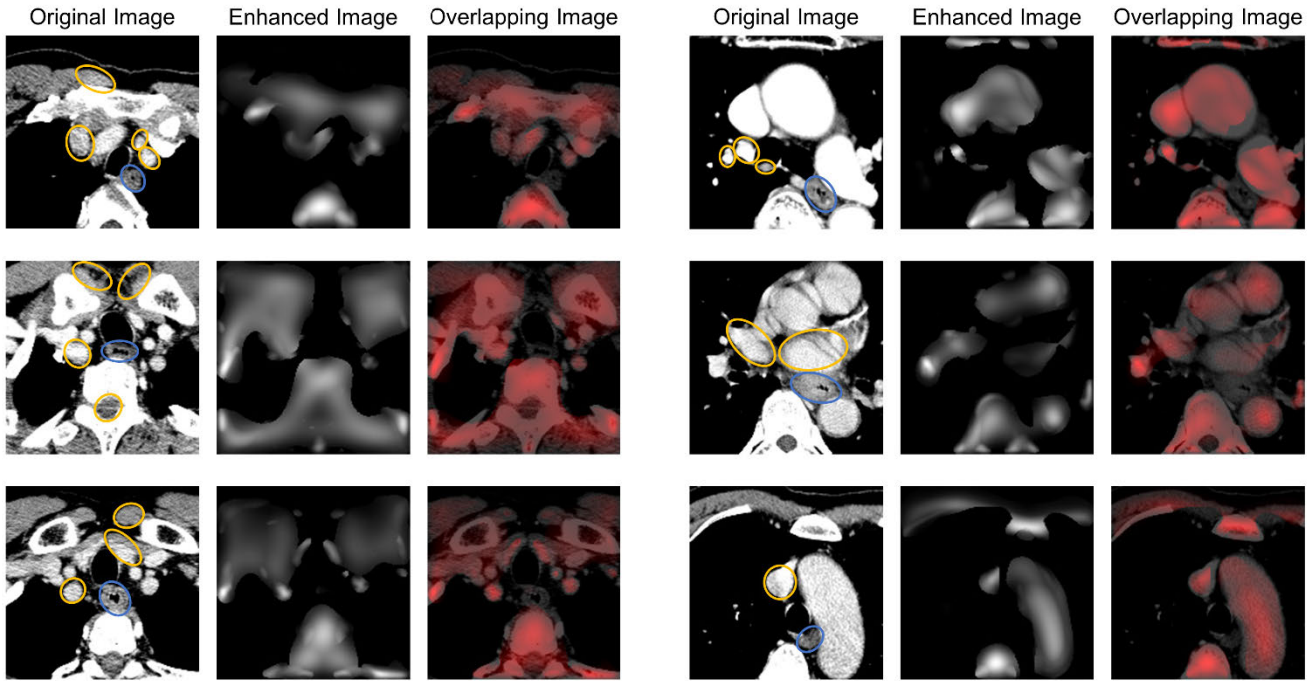


FIGURE 3. Some examples of enhancement results. In the original images, blue ellipses indicate the esophagus and yellow ellipses indicate those regions with similar features to the esophagus, such as size, shape, or texture. In the overlapping images, the enhanced regions are shown in red with different intensities.

Since the structures expected to be enhanced are brighter than the background that is black in CT images, the filter response in all the darker structures is set to 0. R_t , R_p , and R_b denote the similarity measures of the tube-like, plate-like, and blob-like structure, respectively. They are defined as follows:

$$R_t = \frac{|\lambda_2| - |\lambda_1|}{|\lambda_3|} \quad (4)$$

$$R_p = \frac{|\lambda_3| - |\lambda_2|}{|\lambda_3|} \quad (5)$$

$$R_b = \frac{|\lambda_1|}{\sqrt{|\lambda_2 \lambda_3|}} \quad (6)$$

where their values are proportional to the similarity. R_s denotes the similarity measure of evident structures, which is defined as:

$$R_s = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \quad (7)$$

where R_s is small in the background and is large in an evident structure since one of the eigenvalues is large at least. Moreover, α , β , γ , and δ control the sensitivity of MSRF for R_t , R_p , R_b , and R_s , respectively. S_t and S_o denote different scale ranges. S_t matches the size of the tube-like structure except the esophagus, and S_o matches the size of the other structures (plate-like and blob-like structures). Therefore, MSRF can enhance other tissues and organs with evident structures at multiple scales and avoid enhancing the esophagus by setting appropriate S_t and S_o . According to anatomical information of the esophagus and lots of experiments, an optimal parameter combination of MSRF is finally found.

α , β , γ , and δ are set to 0.8, 1, 1.5, and 200. S_t and S_o are set to [5], [10] and [11], [15] respectively, and the step size is set to 1 when choosing scales.

MSRF has non-negative responses between 0 and 1 in each voxel. For each voxel, we integrate the filter responses at multiple scales and take the maximum as the output response:

$$E(x) = \max_{\sigma \in S_o, S_e} E_\sigma(x) \quad (8)$$

Finally, the enhanced 3D volume data is decomposed to consecutive 2D images that are cropped as the input of the segmentation network. The MSRF can enhance the surrounding tissues and organs to extract prior knowledge from anatomical structures, which can assist the segmentation network in channel-wise 2.5D segmentation and strengthen the robustness of the network. Some examples of enhancement results are shown in Fig. 3.

C. NETWORK ARCHITECTURE OF ES0-NET

The basic structure of our 2.5D segmentation network is improved from U-Net [11] that is widely used in medical image segmentation. The symmetrical encoder-decoder architecture with skip connections can extract and combine low-level detail information and high-level semantic information. As shown in Fig. 4, Eso-Net is comprised of different convolutional blocks, and the Prior Attention Module (PAM) is embedded in each connection path between the encoder block and the corresponding decoder block. The concatenated CT images are the input of the network, and the enhanced images are utilized in PAMs. In the last decoder block,

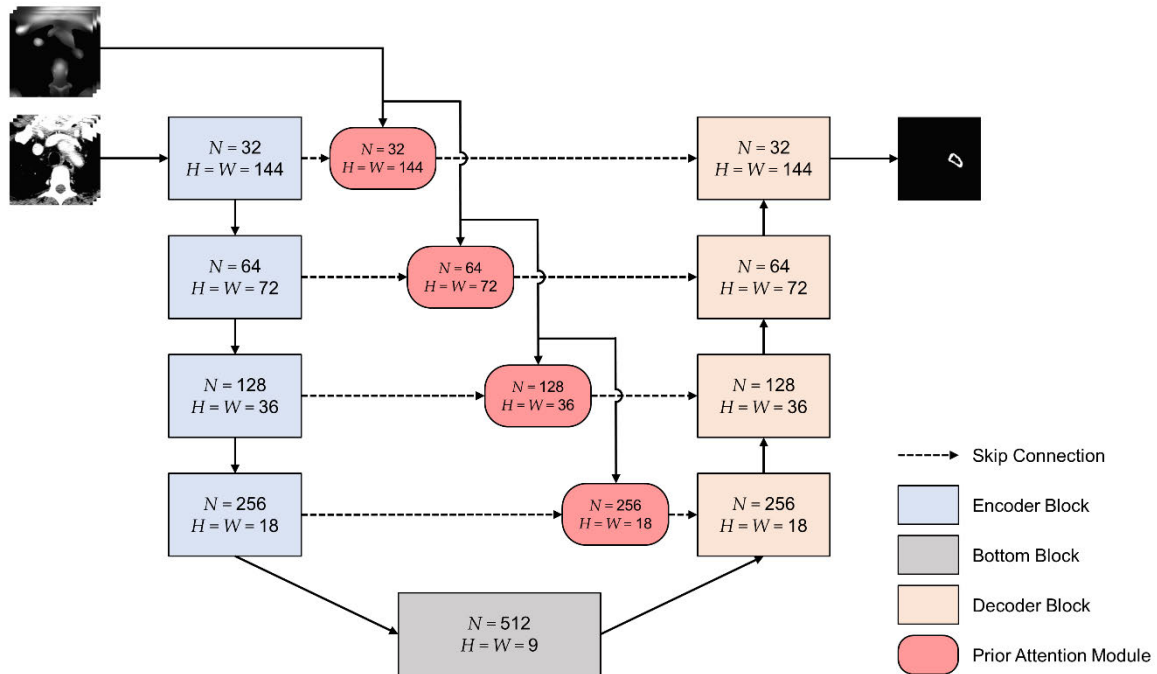


FIGURE 4. The architecture of Eso-Net. H , W , and N represent the height, width and the number of channels of feature maps produced by the convolutional layers, respectively.

a 1×1 convolution is employed at the end to reduce the number of channels to 2, followed by a softmax function to produce two probability maps of the foreground and background. The regions with a higher foreground probability are predicted as the cancerous esophagus. Finally, the segmentation map is generated by the network.

1) CONVOLUTIONAL BLOCKS FOR MULTI-SCALE FEATURE LEARNING

Accurate segmentation of the cancerous esophagus with various scales requires an effective capability of multi-scale feature learning. In the encoding phase, a common approach to extract features of different scales is using standard convolutions in gradually down-sampled feature maps. By contrast, we employ extra dilated convolutions in each encoder block to sufficiently capture contextual information. As shown in Fig. 5(a), a 3×3 convolution is used in the left branch of the encoder block to perform normal feature extraction. In the right branch, three 3×3 dilated convolutions with dilated rates of 2-4 are parallelly employed to extract features in larger receipt fields. Then, the produced feature maps are concatenated, followed by a 1×1 convolution to learn the weights of these extracted features and reduce the number of channels. Next, the feature maps from two branches are added up for feature fusion. Finally, the added-up feature maps are transmitted to the PAM of the same level. Meanwhile, a subsequent max-pooling reduces the size of the added-up feature maps by half.

The bottom block is designed for further feature extraction, which as shown in Fig. 5(b). We employ a 3×3 convolution and four 3×3 dilated convolutions with dilated rates of 2-5 to

capture multi-scale information, and the produced feature maps are added up to aggregate information. In contrast to deepening the network with more down-sampling operations, using the combination of dilated convolutions is an efficient method to achieve multi-scale feature learning with fewer parameters, and avoids reconstructing small objects from excessively small feature maps.

In the decoding phase, the feature maps are gradually restored to the size of the input images. The structure of the decoder blocks is shown in Fig. 5(c). Using deconvolution, the input feature maps are up-sampled by a factor of 2. Then, they are concatenated with the feature maps transmitted from the corresponding PAM, and two consecutive convolutional layers are employed behind. Furthermore, the transmitted feature maps are added to the output of the second convolutional layers. Since the transmitted feature maps contain a wealth of contextual information, it is important to make full use of these feature maps. The residual connection helps the network retain this information and passes it on to the next encoder block or the last 1×1 convolutional layer. Note that each convolutional layer in convolutional blocks is followed by a batch normalization and a rectified linear unit (ReLU) as the activation function unless otherwise stated.

2) PRIOR ATTENTION MODULE

In this task, the segmentation accuracy is limited by the loss of spatial information in the encoding phase. Skip connections can facilitate the recovery of spatial information by concatenating the feature maps from the encoder and decoder blocks. However, not all the activations of the feature maps in skip connections are relevant to the esophagus. To emphasize

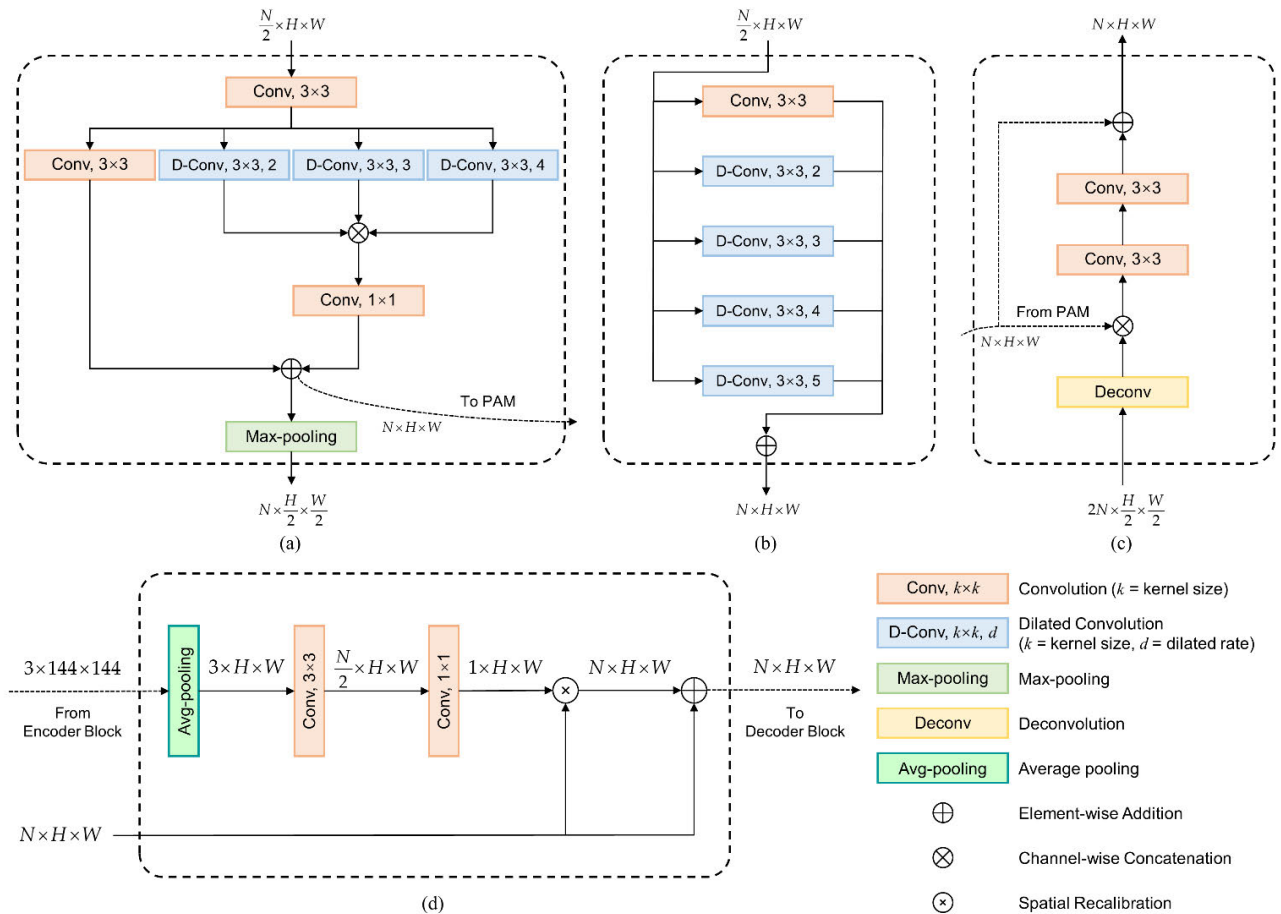


FIGURE 5. The illustrations of convolutional blocks and the Prior Attention Module in Eso-Net: (a) the encoder block; (b) the bottom block; (c) the decoder block; (d) the Prior Attention Module. All convolutions and dilated convolutions are with a stride of 1, and use zero-padding to remain the size of feature maps. The deconvolutions are with a kernel size of 3 and a stride of 2, and use zero-padding to remain the size as well. The max-pooling operations are with a kernel size and a stride of 2. The average-pooling operations are with a kernel size and a stride of $144/H$, where $H = W$. Moreover, H , W , and N have the same meanings as in Figure 4. Note that there is a 1×1 convolution followed by a softmax function at the end of the last decoder block, which is not shown in the illustration for the sake of simplicity.

the esophagus and suppress irrelevant regions, we propose PAM that utilizes the enhanced images to recalibrate the feature maps passed through skip connections. The inputs of PAM include two parts: the enhanced images and the feature map produced by the encoder block of the same level. As shown in Fig. 5(d), the enhanced images are down-sampled to the same size as the input feature maps firstly. Since they cannot be directly considered as the attention map, we employ a 3×3 convolution to capture implicit information of the enhanced images. Then, a 1×1 convolution is employed to reduce the number of channels to 1, followed by a sigmoid function to rescale the activations into $[0,1]$. Hence, the attention map $q \in R^{H \times W}$ is transformed from the enhanced images by two consecutive convolutional layers. Let $M = [m_1, m_2, \dots, m_i, \dots, m_N]$ represents the input feature maps, where $m_i \in R^{H \times W}$ denotes the feature map in channel i , and N is the number of channels. Finally, the output feature maps \hat{M} are given by:

$$\hat{M} = M + [q \odot m_1, q \odot m_2, \dots, q \odot m_i, \dots, q \odot m_N] \quad (9)$$

where \odot denotes an element-wise multiplication. Finally, the recalibrated feature map is transmitted to the decoder block of the same level. With a spatial attention mechanism, the activation of each spatial location is recalibrated by the corresponding scale factor of the attention map. PAMs can learn to emphasize relevant activations and filter irrelevant and noisy activations to make the concatenation operation fuse only useful features. Furthermore, the attention mechanism also works in the back-propagation during training, which allows model parameters in the encoder blocks to be updated mostly based on relevant regions [34].

IV. EXPERIMENTS

A. DATASET DESCRIPTION

The dataset is provided by Sun Yat-sen University Cancer Center and is from 30 esophageal cancer patients with different genders, ages, and cancer stages. The patients include males and females between the ages of 44 and 85. In the dataset, most of the patients are at stage II and III. 14 patients are at stage II and 12 patients are at stage III. Moreover, 1 patient with stage I cancer and 3 patients with stage IV

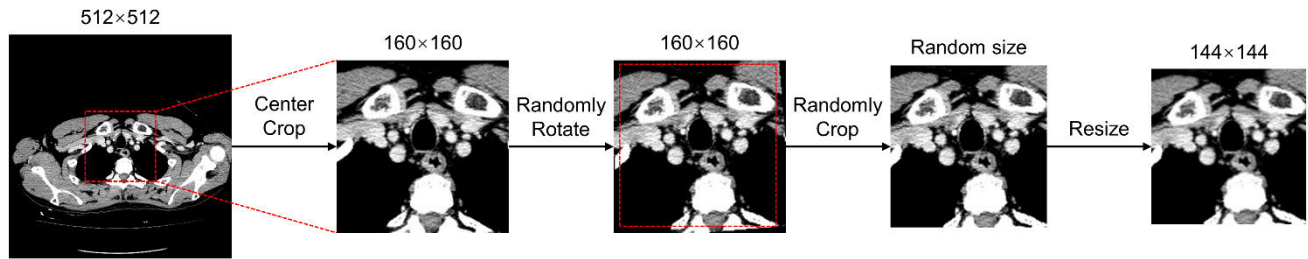


FIGURE 6. The process of our online data augmentation. The image sizes are shown on the top of the images. The operations in different steps are shown above the arrows. The red dotted frames represent the cropped regions.

cancer are also contained in the dataset to increase the data diversity. CT scan of the chest is performed for each patient and then a total of 6362 CT slices are collected. The CT slices have 512×512 pixels in-plane size with 2 mm slice thickness and the pixel spacing varies from 0.625 mm to 0.923 mm in the axial plane. Moreover, all the CT slices are labeled manually by professional doctors.

In dataset division, the CT slices of 5 randomly selected patients are utilized as our test set and the rest are utilized as the training and validation set. Moreover, five-fold cross-validation [39] is performed to evaluate the model training, and 5 models are trained based on different training sets. The CT slices of the remaining 25 patients are randomly split into 5 groups on average so that each group includes 5 patients. In the training of each model, one of the groups is selected as the validation set in turn and the remaining groups are as the training set. We tune the hyper-parameters based on the validation effects of 5 models. Finally, the model is trained on all the CT slices of 5 groups with the optimal hyper-parameter configuration and the performance is evaluated on the test set.

B. IMPLEMENTATION DETAILS

1) DATA AUGMENTATION

Data augmentation is a common approach to enhance model robustness. During training, we perform online data augmentation to increase the diversity of data available for training models and alleviate storage requirements. The process is shown in Fig. 6. Firstly, the 160×160 pixels region in the center of the original 160×160 CT image is cropped. Then, the cropped image is randomly rotated -5° to 5° clockwise and the size remains the same by zero-padding. Next, we randomly crop the rotated image to obtain a random-size image. The aspect ratio of the random-size image is randomly limited within $[0.8, 1.2]$. The area ratio of the random-size image to the rotated image is randomly limited within $[0.85, 0.95]$. Finally, using bilinear interpolation, the random-size image is resized to 144×144 pixels suitable for the segmentation network. Furthermore, we perform the same processing on the enhanced images and the label images.

2) MODEL TRAINING

We using deep learning technology to train the models and then determine the model parameters of the segmentation networks. The adaptive moment estimation (ADAM)

optimizer [40] with a weight decay of 0.002 is employed to realize gradient descent, and the first and second moment estimates are set to 0.9 and 0.999 respectively. We use an initial learning rate of 0.00001 and employ a step-based decay schedule that drops the learning rate by 40% every 20 epochs. The batch size is set to 16 and the model parameters are initialized using the method introduced in [11]. Moreover, we select Dice loss [41] as the loss function. Using five-fold cross-validation, we train the models for 100 epochs and the performance is evaluated on the validation set after each epoch. The best models of different networks are selected for final evaluation on the test set. All the models are implemented with Python 3.7 and PyTorch 1.3.1, and are trained on an NVIDIA GTX 1080TI GPU.

C. EVALUATION METRICS

To measure the performance of the proposed method and other existing methods, dice similarity coefficient (DSC), precision (PRE), sensitivity (SEN) and Hausdorff distance (HD) are employed in our experiments. All of them are usual pixel-level evaluation metrics for biomedical image segmentation. DSC, PRE and SEN can be considered as similarity metrics of two different sets, which are defined as follows:

$$\text{DSC} = \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \quad (10)$$

$$\text{PRE} = \frac{|Y \cap \hat{Y}|}{|\hat{Y}|} \quad (11)$$

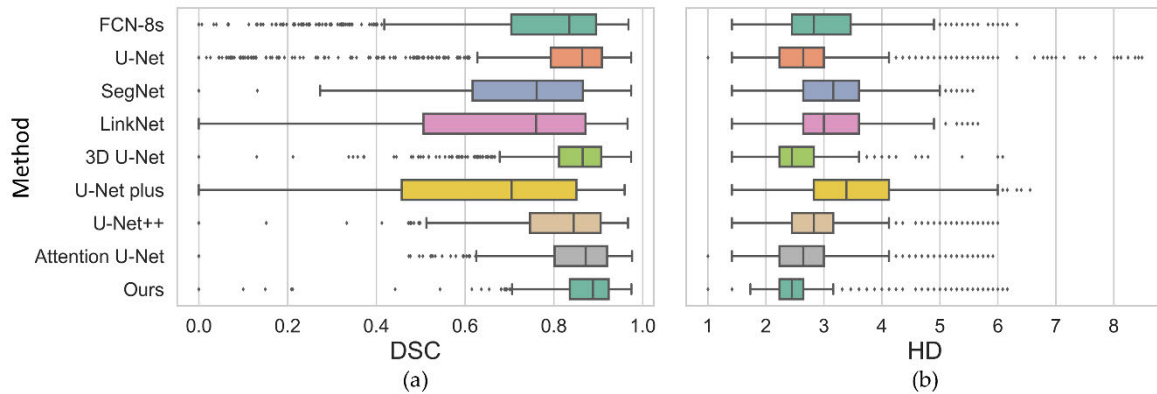
$$\text{SEN} = \frac{|Y \cap \hat{Y}|}{|Y|} \quad (12)$$

where Y denotes the pixel set of the target region (cancerous esophagus) and \hat{Y} denotes the pixel set of the predicted region segmented by the network. $|\cdot|$ means to count the number of pixels in a set. All of them range between $[0, 1]$ and greater DSC, PRE and SEN means better segmentation performance. HD measures the deviation degree of two sets in a metric space, which is defined as:

$$\text{HD} = \max(d(Y, \hat{Y}), d(\hat{Y}, Y)) \quad (13)$$

TABLE 2. Performance Comparisons of the Different Methods.

Method	DSC (%)	PRE (%)	SEN (%)
FCN-8s [8]	75.703	74.479	76.968
U-Net [11]	78.432	80.330	76.622
SegNet [12]	72.167	66.181	79.344
LinkNet [13]	66.073	68.182	64.091
3D U-Net [14]	83.693	84.274	83.120
U-Net plus [15]	61.033	62.713	59.441
U-Net++ [30]	79.025	74.992	83.516
Attention U-Net [34]	82.425	82.494	82.356
Ours	84.839	85.955	83.752

**FIGURE 7.** The box plots of performance of different methods: (a) the box plot of performance in DSC; (b) the box plot of performance in HD.

where $d(Y, \hat{Y})$ and $d(\hat{Y}, Y)$ are defined as follows:

$$d(Y, \hat{Y}) = \max_{y \in Y} (\min_{\hat{y} \in \hat{Y}} \|y - \hat{y}\|) \quad (14)$$

$$d(\hat{Y}, Y) = \max_{\hat{y} \in \hat{Y}} (\min_{y \in Y} \|\hat{y} - y\|) \quad (15)$$

where y and \hat{y} denote the pixel in Y and \hat{Y} respectively. $\|\cdot\|$ means to compute the euclidean distance. HD ranges between 0 and $+\infty$. As HD declines, segmentation performance increases.

D. RESULTS AND ANALYSIS

1) COMPARISONS WITH OTHER METHODS

We compare the proposed method with several segmentation networks including four popular baseline models including FCN-8s [8], U-Net [11], SegNet [12] and LinkNet [13], and four variants of U-Net including 3D U-Net [14], U-Net plus [15], U-Net ++ [30] and Attention U-Net [34]. The performance comparisons are reported in Table 2.

From the experimental results, LinkNet and SegNet achieve poor performance in DSC and HD. One of the reasons is that LinkNet and SegNet are proposed for scene segmentation and are not applicable to segmenting the object with a blurry boundary. Moreover, U-Net and its variants obtain better segmentation results except U-Net plus. The improvement of performance is credited with the correct use of skip connections. Among them, 3D U-Net achieves the best performance with 83.693% DSC, 84.274% PRE,

83.120% SEN, and 2.591mm HD, which reflects the superiority of 3D convolutional networks in medical image segmentation. Two 2D convolutional networks also achieve competitive performance. U-Net++, which integrates U-Nets of different depths, obtains 79.025% DSC, 74.992% PRE, 83.516% SEN, and 2.875mm HD. Attention U-Net, which uses a spatial attention mechanism in skip connections, obtains a higher DSC (82.425%), PRE (82.494%), and a lower HD (2.751mm). Whereas both of them still do not outperform 3D U-Net. The experimental results show that no existing segmentation network can achieve the best performance in all evaluation metrics, since they do not take into account the particularity of the cancerous esophagus. By contrast, the proposed method can tackle this task fairly well. From Table 2, it obtains the highest DSC (84.839%), PRE (85.955%), SEN (83.752%), and the lowest HD (2.583mm), which demonstrates that it outperforms other existing segmentation networks in this task. Moreover, the experiments based on our dataset also verify the strong generalization capacity of the proposed method. As shown in Fig. 7, we draw the box plots of performance in the two most important metrics (DSC and HD) for samples in the test set. We can observe that the proposed method can obtain better performance and relatively steady segmentation accuracy.

The visual examples of segmentation results from the test set are shown in Fig. 8, where each row shows the same CT image. From Fig. 8, we can observe that the size and shape of the esophagus are various. Some segmentation network,

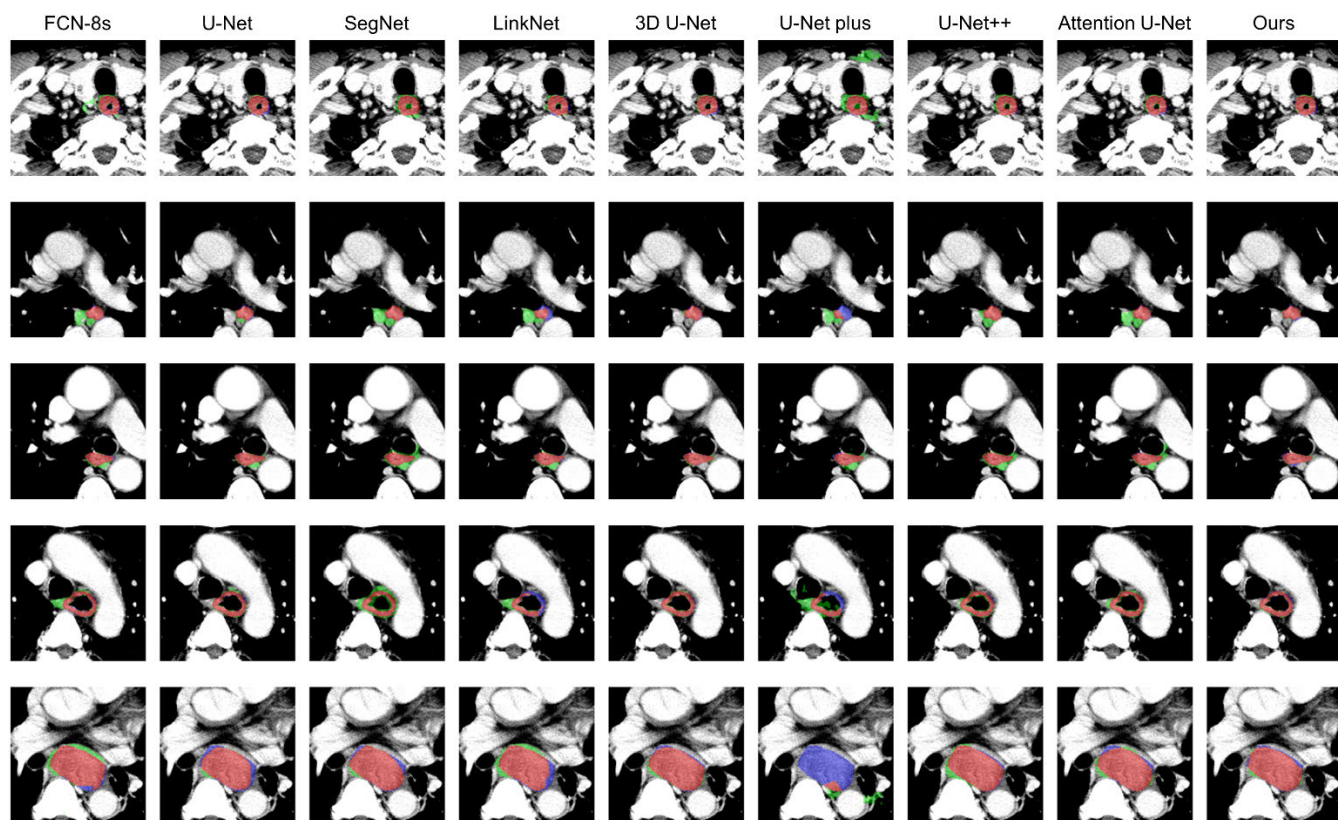


FIGURE 8. The visual examples of segmentation results. The size of CT images is 144×144 . True positives, false positives and false negatives are indicated in red, green, and blue, respectively.

such as FCN-8s, SegNet, LinkNet, and U-Net plus, cannot obtain satisfactory segmentation results. For these networks, some similar tissues and organs are wrongly predicted as the cancerous esophagus. Since the region of the cancerous esophagus in CT images is the key diagnostic basis, inaccurate segmentation results can bring serious consequences. By contrast, the proposed method can obtain high accurate segmentation results. The last row in Fig. 8 shows a challenging case for segmentation, where the esophagus with a huge tumor has a blurry boundary and low contrast against surrounding tissues. In this case, the proposed method achieves better segmentation performance than the other networks. These visual examples show that the proposed methods can provide reliable segmentation results for the diagnosis and treatment of esophageal cancer.

2) ABLATION STUDY

We conduct an ablation study to show the effectiveness of each improvement in the proposed method. The experimental results are shown in Table 3. For simplicity, we use Model 1-4 to represent the models in Table 3 from top to bottom. Model 1, which is indicated by “BS” in Table 3, accepts a single CT image as input. It outperforms the original U-Net in DSC, SEN, and HD. By contrast, Model 2 requires three CT images as input to perform channel-wise 2.5D segmentation that brings 1.828% DSC and 0.048mm HD improvement.

However, Model 2 still perform poorly than some existing networks such as Attention U-Net and 3D U-Net. Model 3 embeds PAMs in skip connections and utilizes the original CT images instead of the enhanced images to generate attention maps. From Table 3, PAMs can bring 1.121% DSC improvement due to the use of a spatial attention mechanism in skip connections. At last, Model 4 represents the proposed method. The experimental result of Model 4 shows that MSRF brings a relatively huge improvement in PRE and HD. This is because the enhanced images produced by MSRF can guide the network to distinguish irrelevant tissues and organs, which contributes to eliminate most false positives.

The convergence curves of Mode 1-4 are shown in Fig. 9. Even with trainable parameters increased, the models with PAMs have higher convergence efficiency than those models without PAMs. The major reason is that PAMs can make the encoder blocks of the network mostly focus on the emphasized region during the back-propagation. Moreover, MSRF further facilitates the model convergence since the enhanced images produced by it can provide additional prior knowledge for the network.

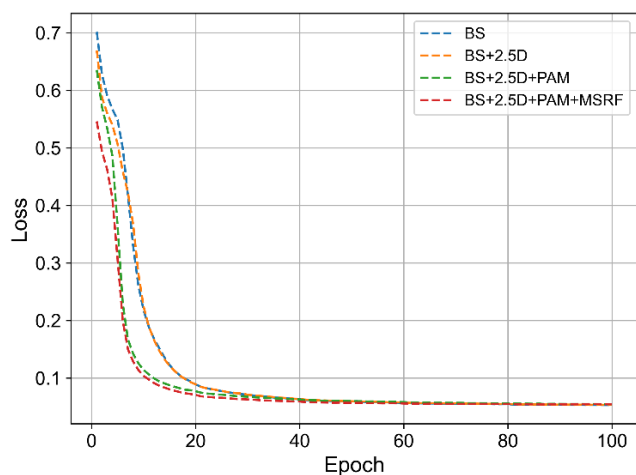
V. DISCUSSION

The various anatomical structure and the blurry boundary of the cancerous esophagus are the key problems limiting the improvement of segmentation accuracy. Currently,

TABLE 3. Quantitative comparisons of each improvement in the proposed method.

Method	DSC (%)	PRE (%)	SEN (%)
BS (Model 1)	79.628	78.579	80.705
BS+2.5D (Model 2)	81.456	81.661	81.252
BS+2.5D+PAM (Model 3)	82.577	81.651	83.524
BS+2.5D+PAM+MSRF (Model 4)	84.839	85.955	83.752

“BS” indicates the basic structure of Eso-Net without PAM in skip connections. “2.5D” indicates that the network performs channel-wise 2.5D segmentation. “PAM” and “MSRF” indicate the Prior Attention Module and the Multi-Structure Response Filter respectively.

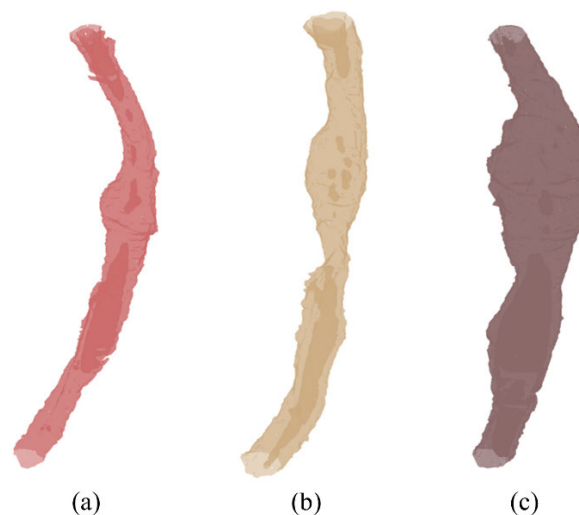
**FIGURE 9.** The convergence curves of Model 1-4. The loss in the y-axis represents the average loss per sample during training.

the performance of existing segmentation networks is not satisfactory in this task. In this study, we have proposed a novel segmentation network named Eso-Net and a 3D enhancement filter called Multi-Structure Response Filter (MSRF) to address these problems. The experiments demonstrate that the proposed method outperforms existing segmentation networks in this task.

The experimental results show that channel-wise 2.5D segmentation is conducive to improve the performance, since it can efficiently utilize z-axis information with fewer parameters. Moreover, our experiments demonstrate that using only standard convolutions for multi-scale feature extraction is not enough to achieve highly accurate segmentation results in this task. By contrast, we parallelly use multiple dilated convolutions in the same feature maps, and residual connections are employed for feature fusion and facilitating gradient propagation. The validity of this approach is verified by our experiments. Moreover, MSRF was designed to enhance the regions of irrelevant tissues and organs, and the enhanced images are utilized in PAMs. The experimental results demonstrate that the use of MSRF and PAMs takes a positive effect on the improvement of the segmentation performance.

Generally, Doctors observe CT slices one by one to diagnose the stage and location of esophageal cancer. However, it is not convenient for diagnosis and clinical analysis due to heavy workload and tiresome procedures. Automatic

segmentation based on deep learning can effectively assist doctors. The proposed method can take the place of doctors to accomplish this tedious and time-consuming work, since it is automatic and achieves state-of-the-art performance in this task. Verified by professional doctors, the segmentation results of the proposed method achieve the accuracy needed in practical clinical applications. As shown in Fig. 10, the output segmentation maps can be used to generate a 3D model that presents the entire esophagus of a patient. Doctors can conveniently observe the shape and structure of the tumor at the esophagus for further diagnosis and treatment. Moreover, it is easier to measure the size of the esophageal tumor on a 3D model than on 2D CT images.

**FIGURE 10.** The 3D model generated from our segmentation results. (a), (b), and (c) show esophageal 3D models of different patients in the test set.

However, the proposed method also has its limitations. Firstly, the generalization capacity of Eso-Net is limited by the small size of the cancerous esophagus dataset. In this case, Eso-Net may obtain poor results when segmenting some extremely special samples. Secondly, MSRF is an untrainable component independent of the network. When applying the proposed method in other medical image tasks, we need to tune its parameters manually before the model training. Hence, our future goals are optimizing the architecture of Eso-Net on a larger dataset and merging MSRF with the deep

learning model, which can further facilitate the development of medical image segmentation.

VI. CONCLUSION

In this article, we have proposed a novel 2.5D segmentation network named Eso-Net for automatic segmentation of the cancerous esophagus and designed a 3D enhancement filter called Multi-Structure Response Filter (MSRF) to extract 3D structural priors. Eso-Net is based on an encoder-decoder architecture and consists of different convolutional blocks. Dilated convolutions and residual connections are employed in the convolutional blocks to facilitate multi-scale feature extraction and fusion. Furthermore, the proposed Prior Attention Modules (PAM) are embedded in skip connections to recalibrate the activations of feature maps with the assistance of the enhanced images. In the experiments, the proposed method reports the highest DSC (84.839%), PRE (85.955%), SEN (83.752%), and the lowest HD (2.583mm), which demonstrates the proposed method achieve the best performance in automatic segmentation of the cancerous esophagus. Moreover, the ablation study shows that each improvement of the proposed method contributes to obtain better segmentation performance. In the future, we are interested in optimizing the architecture of Eso-Net on a larger dataset and merging MSRF with the deep learning model.

REFERENCES

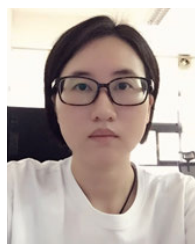
- [1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.
- [2] B. Gupta and N. Kumar, "Worldwide incidence, mortality and time trends for cancer of the oesophagus," *Eur. J. Cancer Prevention*, vol. 26, no. 2, pp. 107–118, Mar. 2017.
- [3] K. M. Jang, K. S. Lee, S. J. Lee, E. A. Kim, T. S. Kim, D. Han, and Y. M. Shim, "The spectrum of benign esophageal lesions: Imaging findings," *Korean J. Radiol.*, vol. 3, no. 3, pp. 199–210, 2002.
- [4] M. Rousson, Y. Bai, C. Xu, and F. Sauer, "Probabilistic minimal path for automated esophagus segmentation," in *Proc. Med. Imag. Image Process.*, Mar. 2006, Art. no. 614449.
- [5] J. Feulner, S. K. Zhou, M. Hammon, S. Seifert, M. Huber, D. Comanicu, J. Hornegger, and A. Cavallaro, "A probabilistic model for automatic segmentation of the esophagus in 3-D CT scans," *IEEE Trans. Med. Imag.*, vol. 30, no. 6, pp. 1252–1264, Jun. 2011.
- [6] D. Grosgeorge, C. Petitjean, B. Dubray, and S. Ruan, "Esophagus segmentation from 3D CT data using skeleton prior-based graph cut," *Comput. Math. Methods Med.*, vol. 2013, pp. 1–6, Jul. 2013.
- [7] J. Yang, B. Haas, R. Fang, B. M. Beadle, A. S. Garden, Z. Liao, L. Zhang, and P. Balter, "Atlas ranking and selection for automatic segmentation of the esophagus from CT scans," *Phys. Med. Biol.*, vol. 62, no. 23, p. 9140, 2017.
- [8] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
- [9] T. Fechter, S. Adebahr, D. Baltas, I. Ben Ayed, C. Desrosiers, and J. Dolz, "Esophagus segmentation in CT via 3D fully convolutional neural network and random walk," *Med. Phys.*, vol. 44, no. 12, pp. 6341–6352, Dec. 2017.
- [10] R. Trullo, C. Petitjean, D. Nie, D. Shen, and S. Ruan, "Fully automated esophagus segmentation with a hierarchical deep learning approach," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 503–506.
- [11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Munich, Germany, 2015, pp. 234–241.
- [12] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [13] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
- [14] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense, volumetric segmentation from sparse annotation," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Athens, Greece, 2016, pp. 424–432.
- [15] S. Chen, H. Yang, J. Fu, W. Mei, S. Ren, Y. Liu, Z. Zhu, L. Liu, H. Li, and H. Chen, "U-Net plus: Deep semantic segmentation for esophagus and esophageal cancer in computed tomography images," *IEEE Access*, vol. 7, pp. 82867–82877, 2019.
- [16] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multi-scale vessel enhancement filtering," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Cambridge, MA, USA, 1998, pp. 130–137.
- [17] Y. Sato, C. Westin, A. Bhalerao, S. Nakajima, N. Shiraga, S. Tamura, and R. Kikinis, "Tissue classification based on 3D local intensity structures for volume rendering," *IEEE Trans. Vis. Comput. Graphics*, vol. 6, no. 2, pp. 160–180, Apr. 2000.
- [18] Q. Li, S. Sone, and K. Doi, "Selective enhancement filters for nodules, vessels, and airway walls in two- and three-dimensional CT scans," *Med. Phys.*, vol. 30, no. 8, pp. 2040–2051, Jul. 2003.
- [19] D. Jimenez-Carretero, A. Santos, S. Kerkstra, R. D. Rudyanto, and M. J. Ledesma-Carbayo, "3D frangi-based lung vessel enhancement filter penalizing airways," in *Proc. IEEE 10th Int. Symp. Biomed. Imag.*, Apr. 2013, pp. 926–929.
- [20] A. Shahzad, C. M. Goh, N. M. Saad, N. Walter, A. S. Malik, and F. Meriaudeau, "Subcutaneous veins detection and backprojection method using frangi vesselness filter," in *Proc. IEEE Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, Apr. 2015, pp. 65–68.
- [21] M. Shahid and I. A. Taj, "Robust retinal vessel segmentation using vessel's location map and Frangi enhancement filter," *IET Image Process.*, vol. 12, no. 4, pp. 494–501, Apr. 2018.
- [22] H. Jiang, H. Ma, W. Qian, M. Gao, and Y. Li, "An automatic detection system of lung nodule based on multigroup patch-based deep learning network," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 4, pp. 1227–1237, Jul. 2018.
- [23] A. G. Blaiech, A. Mansour, A. Kerkeni, M. H. Bedoui, and A. B. Abdallah, "Impact of enhancement for coronary artery segmentation based on deep learning neural network," in *Proc. Pattern Recognit. Image Anal.*, Madrid, Spain, 2019, pp. 260–272.
- [24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: <http://arxiv.org/abs/1511.07122>
- [25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.
- [26] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.
- [27] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1451–1460.
- [28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
- [29] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 405–420.
- [30] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
- [31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
- [32] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 3–19.

- [33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.
- [34] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*. [Online]. Available: <http://arxiv.org/abs/1804.03999>
- [35] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Proc. Med. Image Comput. Comput. Assist. Interv. (MICCAI)*, Granada, Spain, 2018, pp. 421–429.
- [36] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-Net for biomedical image segmentation," 2018, *arXiv:1812.04103*. [Online]. Available: <http://arxiv.org/abs/1812.04103>
- [37] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink, and M. A. Viergever, "Scale and the differential structure of images," *Image Vis. Comput.*, vol. 10, no. 6, pp. 376–388, Jul. 1992.
- [38] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *Int. J. Comput. Vis.*, vol. 30, no. 2, pp. 117–156, 1998.
- [39] M. Stone, "Cross-validator choice and assessment of statistical predictions," *J. Roy. Stat. Soc., Ser. B*, vol. 36, no. 2, pp. 111–133, 1974.
- [40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: <http://arxiv.org/abs/1412.6980>
- [41] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.



oncologic imaging and biomedical analysis.

SIYU ZHU received the B.Sc. and M.Sc. degrees from Sun Yat-sen University, in 2011 and 2014, respectively, and the Ph.D. degree in biomedical science from The University of Hong Kong, in 2018. She is currently a Radiologist with the State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China. Her research interests include



ZHUOWEI WANG received the Ph.D. degree in computer system architecture from Wuhan University, Wuhan, China, in 2012. She is currently an Associate Professor with the Institute of Computing, Guangdong University of Technology. Her research interests include high-performance computing, low-power optimization, and distributed systems.



DONGHAO ZHOU is currently pursuing the bachelor's degree in information engineering with the School of Information Engineering, Guangdong University of Technology, Guangzhou, China. He is also a member of the Guangdong Key Laboratory of Cyber-Physical System and hosts an Innovation And Entrepreneurship Project for college students. His research interests include image segmentation and medical image analysis.



GUOHENG HUANG (Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics and the M.Eng. degree in computer science from South China Normal University, in 2008 and 2012, respectively, and the Ph.D. degree in software engineering from the University of Macau, in 2017. He is currently an Assistant Professor of computer science with the Guangdong University of Technology. His research interests include computer vision, pattern recognition, and artificial intelligence. He is currently a CCF Member. He has hosted and undertaken a number of national and provincial-level scientific research projects, including the Natural Science Foundation of China and the National Key Research and Development Plan. As a Key Member of the Guangdong Key Laboratory of Cyber-Physical System, he has published many research articles.



JIAJIAN LI was born in Maoming, Guangdong, China, in 1995. He received the B.E. degree in computer science and technology from Zhaoqing University, Zhaoqing, China, in 2017. He is currently pursuing the M.S. degree in computer technology with the Guangdong University of Technology. His research interests include image processing of medical image detection and deep learning.



BINGO WING-KUEN LING (Senior Member, IEEE) received the B.Eng. (Hons.) and M.Phil. degrees from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, in 1997 and 2000, respectively, and the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, in 2003. In 2004, he joined the King's College London as a Lecturer. In 2010, he joined the University of Lincoln as a Principal Lecturer, where he was promoted to a Reader in 2011. In 2012, he joined the Guangdong University of Technology as a Full Professor. He has published an undergraduate textbook, a research monograph, several book chapters, a book review published in an IEEE journal, more than 150 internationally leading journal articles, and more than 130 highly rated international conference papers. His research interests include the time frequency analysis, the optimization theory, the symbolic dynamics, the multimedia signal processing, and the biomedical signal processing. He is a Fellow of IET, a China National Young Thousand-People-Plan Distinguished Professor, and a University Hundred-People-Plan Distinguished Professor. He was awarded the Best Reviewer Prizes from the IEEE Instrumentation and Measurement Society, in 2008 and 2012. He serves in the technical committees of the Nonlinear Circuits and Systems Group, the Digital Signal Processing Group, and the Power Electronics and Systems Group of the IEEE Circuits and Systems Community. He has also served as the Guest Editor-in-Chief for several special issues of highly rated international journals, such as *Circuits, Systems, and Signal Processing* and the *American Journal of Engineering and Applied Sciences*. He is an Associate Editor of *Circuits, Systems, and Signal Processing*, the *Journal of the Franklin Institute*, *Measurement*, *IET Signal Processing*, and the *Journal of Industrial and Management Optimization*.



CHI-MAN PUN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau, in 1995 and 1998, respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, in 2002. He is currently an Associate Professor and the Head of the Department of Computer and Information Science, University of Macau. He has investigated several funded research projects and authored or coauthored more than 100 refereed scientific papers in international journals, books, and conference proceedings. He has also served as the Editorial Member/Referee for many international journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. His research interests include digital image processing, multimedia forensics and watermarking, pattern recognition, and computer vision. He is also a Professional Member of ACM.



LIANGLUN CHENG (Senior Member, IEEE) received the B.E. and M.S. degrees in automation from the Huazhong University of Science and Technology, Wuhan, China, in 1988 and 1992, respectively, and the Ph.D. degree in automation from the Chinese Academy of Sciences, in 1999. He is currently a Professor and the Computer Dean with the Guangdong University of Technology, a Doctoral Tutor, an Excellent Teacher of Nanyue, and a National-Level Target Trainer for the Thousand-Ten Thousand Project of cross-century talents in Guangdong Province. He is also the Executive Director of the Robotics Professional Committee of the China Automation Association, a member of the China Computer Federation, and the Vice Chairman of the Guangdong Automation Association. His main research interests include knowledge graph, knowledge automation, and information physics fusion systems.



XIUYU CAI is currently a Professor of medical oncology, an Associate Senior Doctor, an Associate Professor, and a master's Tutor with the Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China. He has published many research articles as the first/co-first/ corresponding author. His research interests include immunotherapy, targeted therapy, and comprehensive treatments of common tumors, especially in lung cancer, nasopharyngeal carcinoma, esophageal carcinoma, hepatocellular carcinoma, colorectal cancer, and breast cancer.



JIAN ZHOU is currently a Radiologist with the State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Department of Medical Imaging, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China. His research interest includes data mining in medical.

...