IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 2, 2020, accepted July 1, 2020, date of publication July 7, 2020, date of current version July 17, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007719

Channel-Attention U-Net: Channel Attention Mechanism for Semantic Segmentation of Esophagus and Esophageal Cancer

GUOHENG HUANG[1], (Member, IEEE), JUNWEN ZHU[1], JIAJIAN LI[1], ZHUOWEI WANG[1], LIANGLUN CHENG[1], (Senior Member, IEEE), LIZHI LIU[2], HAOJIANG LI[2], AND JIAN ZHOU[2]

[1]School of Computers, Guangdong University of Technology, Guangzhou 510006, China
[2]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 510060, China

Corresponding authors: Haojiang Li (lihaoj@sysucc.org.cn) and Jian Zhou (zhoujian@sysucc.org.cn)

This work was supported in part by the Guangdong Esophageal Cancer Institute Science and Technology Program under Grant Q-201602, in part by the National Natural Science Foundation of China under Grant 61702111, in part by the National Natural Science Foundation of Guangdong Joint Fund under Grant U1801263 and Grant U1701262, in part by the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2016B030301008, in part by the Guangdong Research and Development Plan Projects in Key Areas under Grant 2018B010109007, in part by the Key-Area Research and Development Program of Guangdong Province under 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 The effective segmentation of esophagus and esophageal cancer from Computed Tomography (CT) images can meaningfully assist doctors in the diagnosis and treatment of esophageal cancer patients. However, problems such as the small proportion of the esophageal region in CT images and the irregular shape of the esophagus will make the diagnosis difficult. In practical applications, not all esophagus and esophageal cancer morphology can be included in the training set, so the generalization ability of the model is very important. These are the difficulties in segmenting the esophagus and esophageal cancer. Since some adjacent tissues and organs of the esophagus are visually close to the esophagus and esophageal cancer, how to ensure that the network can extract effective distinguishing features has become the focus of research. In this paper, a novel U-Net structure — Channel-attention U-Net is proposed to segment esophagus and esophageal cancer from CT slices. This novel network combines a Channel Attention Module (CAM) that can distinguish the esophagus and surrounding tissues by emphasizing and inhibiting channel feature and Cross-level Feature Fusion Module (CFFM) which is utilized to strengthen the generalization ability of the network by using high-level features to weight low-level features. Because the high-level features represent specific organizational information, and the low-level features represent the characteristics of detailed information such as edges and contours, the network can learn specific detailed features of a definite organization. In addition, to locate the esophageal region better, a 3D semi-automatic method for segmenting esophagus and esophageal cancer is proposed. The proposed network is trained using 46,400 CT pictures as the training set and divides 11,600 CT images from the dataset at a ratio of 0.2 as the validation set. Finally, 7,250 CT images were used as the test set to test the performance of the network. The experimental results show that the IoU value of our network can reach 0.625, the dice value is 0.732 and the Hausdorff distance is 3.193.

INDEX TERMS Esophageal cancer, channel attention mechanism, deep learning, computed tomography (CT).

I. INTRODUCTION

Esophageal cancer is one of the most common cancers worldwide, and their incidence has been increasing in recent years [1]. For cancer, early diagnosis and treatment are the keys to improving survival rates, and medical imaging technology has provided great help to this. Among many different imaging methods, Computed Tomography (CT) images are widely used for the diagnosis of esophagus diseases because they can provide relatively high-resolution anatomical

The associate editor coordinating the review of this manuscript and approving it for publication was Anandakumar Haldorai.

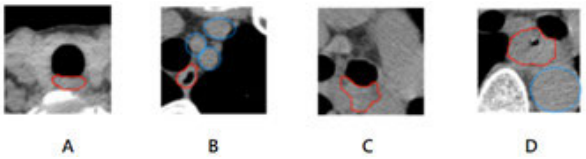


FIGURE 1. Esophagus and esophageal cancer in CT scans. The red curves are the borders between the normal esophagus and the esophageal cancer. The blue curves are the borders of area with similar appearance features to the esophagus or esophageal cancer. A is the normal esophagus without air hole; B is the normal esophagus with air hole; C is the normal esophagus tumor without air holes; D is the normal esophagus tumor with air holes.

information. However, in clinical practice, manual segmentation not only has a large amount of data and is time-consuming and laborious, and different physicians have different diagnoses and segmentation of the same case. Moreover, most tumors have low contrast with surrounding normal tissue, and the borders are blurred. As shown in FIGURE 1, even professional doctors can not accurately point out the areas of the esophagus and esophageal cancer. Because of the need for accurate and effective tumor mapping, the development of semi-automatic or automatic tumor accurate methods is particularly urgent.

A lot of methods have been proposed to extract esophagus contours using traditional image processing methods. Rousson *et al.* introduced a probabilistic shortest path approach to extract the esophagus from CT images [2]. Because the left atrium (LA) and aorta are candidates for esophagus, this method manually places two points on the centerline, and segmentation of the left atrium and aorta as input. In order to extract a new centerline, a probability criterion is defined from the Bayesian formula that combines prior information with image data. Finally, relying on the extracted centerline, the coupled ellipse fitting can reliably detect the outer boundary of the esophagus. Faulkner *et al.* proposed a two-step method that first found the approximate shape using a ''detects and connects'' approach [3]. A classifier is trained to find short segments of the esophagus. Recently developed techniques in discriminative learning and pruning of the search space enable rapid detection of possible esophagus segments. Prior shape knowledge of the complete esophagus is modeled using a Markov chain framework. Yang *et al.* proposed a method which was an online atlas selection approach to select a subset of optimal atlases for multi-atlas segmentation to the delineate esophagus automatically [4]. Damien *et al.* introduced an original method to segment the 3D esophagus from thoracic CT scans using a skeleton-shape model to guide the segmentation [5].

However, in traditional methods, feature extraction mainly relies on artificially designed extractors and requires complex parameter adjustment processes. At the same time, each method is specific to the specific application, and its generalization ability and robustness are not good. With the rise of deep learning, the feature extraction structure is proposed and widely used in the field of medical image. For the segmentation of esophagus and esophageal cancer, the Convolution Neural Network (CNN) is trained to extract the features of the esophagus and esophageal cancer in CT images, and then the probability map of the esophagus and esophageal cancer is obtained by softmax function [6]. Then CNN-based semantic segmentation networks such as Fully Convolutional Network (FCN), SegNet and U-Net were proposed. Hao *et al.* applied the FCN as a classifier for feature extraction of esophageal tumors and successfully segment the area of esophageal tumors [7]. Moreover, SegNet is used in the segmentation of the lung field and cross-sectional brain MRI [8], [9]. However, the networks of encoder-decoder architecture such as FCN and SegNet still have its shortcomings. For example, in the up-sampling phase, the feature extraction part is combined only during the last up-sampling. That will cause the network to lack the ability to recognize small objects due to insufficient utilization of multi-scale feature information when the network is up-sampling. To tackle the above issues, U-Net was proposed [10]. Chen *et al.* proposed U-Net plus which combines two U-shaped structures as feature extractors to segment the esophagus [11]. It has a large number of feature channels, which allow the network to propagate context information to higher resolution layers in the up-sampling part and consider using the concatenate operation instead of element-wise addition operation in skip connection. To introduce the attention mechanism to U-Net, Attention U-Net embeds the attention mechanism into U-Net's skip connection [12]. By extracting a heat map of the shallow features, it can pass the heat map to the decoder to emphasize the relevant areas or suppress the unrelated areas. However, since the appearance of the esophagus and esophageal cancer is various, the network not only needs to segment the esophagus that has appeared in the training set but also needs to segment the esophagus that does not appear in the training set to some extent. This puts higher requirements on the generalization ability of U-Net.

First, it can be seen from the feature extraction networks (such as FPN, U-Net, and SegNet) that the interaction between high-level semantic features and low-level semantic features is important [13]. Low-level extracted features are only some features such as contours, locations and textures, etc. The features extracted with the deepening of the network may be higher semantic features such as esophagus and trachea etc. Existing methods are not effective enough in segmenting the esophagus with blurred boundaries [14]–[17]. Besides, one of the reasons why the existing methods are not effective enough in segmenting the esophagus is that the shape of the esophagus or esophageal cancer is uncertain. In the feature extraction stage of semantic segmentation, the information represented by low-level features and high-level features is different. Low-level features represent information such as texture, boundaries, and contours, while high-level features represent semantic information. Therefore, in the up-sampling phase, the fusion of high-level features and low-level features can promote the segmentation effect of the network.

To achieve the above-mentioned problem, we propose a Cross-level Feature Fusion Module (CFFM) that can gradually fusion features from high to low. Through CFFM, the network can learn the shape and contour features of the esophagus or esophageal cancer to a certain extent. Although feature extraction can obtain the information of the picture, not all information is needed, so we propose a channel attention module (CAM) to filter the features extracted from the picture. CAM can use high-level features to guide the selection of low-level features, thereby establishing the relationship between high-level features and low-level features. By embedding CAM in CFFM, CFFM can select specific features to be integrated into low-level features. Therefore, the network can learn the general shape of the esophagus, which helps to improve the network segmentation effect of the esophagus.

In summary, the main contributions of our work are as follows:

1 The Channel Attention Module (CAM) is adopted to improve the sensitivity of the model to different tissues in the same CT image. This makes it easier for the network to segment ambiguous boundaries between different organizations.

2 A novel model named Cross-level Feature Fusion Module (CFFM) is proposed. On the one hand, CFFM improves the ability of the network to recognize small targets by combining information on multiple scales. On the other hand, CFFM enables each high-level feature to guide the selection of lower-level features. Therefore, the network can learn the shape and contour of the esophagus to improve the generalization ability of the network.

3 Because the esophagus is a narrow coronal structure, it is more difficult to diagnose from a single CT section. Rather, the diagnosis needs to be made in conjunction with a reconstructed 3D model (If a certain part of the esophagus is swollen in a three-dimensional model, it's possible that the diseased part). Therefore, rapid segmentation as well as three-dimensional reconstruction is very important. Due to the small proportion of esophageal and esophageal cancer in the entire CT slice image, we adjusted the size of the original CT image from 512×512 to 80×80 and then 3D rendered the esophageal and esophageal cancers. On the one hand, this allows the separation of esophageal and esophageal cancer at a better image scale. On the other hand, the results of the 2D segmentation will be entered into the proposed 3D semi-automatic segmentation method for 3D esophageal and esophageal cancer segmentation. This allows for rapid 3D esophageal segmentation while ensuring the accuracy of segmentation. In the three major indicators Intersection over Union (IoU), Dice Value (DV) and Hausdorff Distance (HD), we have outperformed over the state-of-the-art methods. To summarize, our approach was first performed by segmenting the 2D CT slices, followed by 3D rendering to reconstruct the 3D model of the esophagus and esophageal cancer. Finally, doctors can make a secondary diagnosis based on this model and verify the correctness of the algorithm's results.

The rest of this article is organized as follows. First of all, we review the techniques mostly related to the proposed framework in Section II. The Section III introduces the preprocessing process of the dataset and details of our network and its components. To verify the effectiveness of our method, extensive experiments are performed in Section IV and then summarize in Section V.

II. RELATED WORKS

A. U-NET AND ITS VARIANTS OF SEMANTIC SEGMENTATION TASKS IN MEDICAL IMAGES

Instead of directly supervising and loss back-propagating on high-level semantic features, U-Net uses the skip connection in the same state because of its symmetrical structure. It also enables the integration of features of different scales to enable multi-scale prediction. Therefore, U-Net is typically used in the field of medical imaging. In order to train deeper networks and extract deep semantic information, Alom *et al.* proposed R2U-Net to combine Recurrent Residual Convolutional Neural Network (RRCNN) with U-Net [18]. R2U-Net achieves good results in vascular, lung and skin datasets. Due to different tasks, there are different requirements for the depth of the U-Net network. Therefore, Zhou *et al.* proposed U-Net++, which embedded U-Net of different scale levels in U-Net, and added deep supervision to allow the network to choose the depth suitable for itself through training [19]. Res U-Net is inspired by residual connections, replacing each sub-module of U-Net with a form with residual connections [20]. Xiao *et al.* applied Res U-Net to segment retinal images and achieved good results. In U-Net, the shallow features are passed to the decoder via skip connection are not filtered, resulting in the network not emphasizing or suppressing certain features during training and affecting the accuracy. Attention U-Net solves the problem that low-level features are not filtered when they are passed through a skip connection by introducing a spatial attention mechanism. However, sometimes the border between esophagus and the esophageal cancer is very blurred. Therefore, the spatial attention models such as Attention U-Net that focus on specific locations of feature maps are not very suitable for segmenting the esophagus and esophageal cancer. Therefore, we need to introduce other attention mechanisms under the premise of U-Net as the basic framework.

B. ATTENTION MECHANISM IN IMAGE FEATURE ENHANCEMENT

The attention modules have been widely used in various fields of deep learning in recent years. Various attention mechanisms have been widely used in feature enhancement of image processing and natural language processing [21]–[23]. Fu *et al.* applied both the spatial attention mechanism and the channel attention mechanism in the semantic segmentation task to make the network to learn the correlation of spatial features and the correlation of modeling channel [24]. Oktay *et al.* presented Attention U-Net to segment pancreas. But Attention U-Net has not applied the channel attention

mechanism to strengthen the connection between the features of the channels. Hu *et al.* presented SE-Net which automatically learns the importance of each feature channel and then uses this importance to enhance useful features and suppress useless features [25]. In order to solve the problem that the network does not have high ability to segment ambiguous boundaries between different organizations in CT images, we explicitly model the interdependence between feature channels in U-Net, and design Channel-attention U-Net which embed a Channel Attention Module (CAM) in the skip connection layer of U-Net.

C. CROSS-LEVEL FEATURE FUSION

In deep learning-based semantic segmentation tasks, U-Net has gradually been the baseline structure of feature extraction. U-Net uses skip connection to connect the feature information of encoder and decoder. Skip connection can effectively fuse the feature information of the down-sampling stage and up-sampling stage. Moreover, U-Net has a symmetrical encoder-decoder structure. Therefore, there are many network structures based on U-Net in the field of medical image [26]–[30]. Zhou *et al.* introduced U-Net++ that the architecture was essentially a deeply-supervised encoder-decoder network where the encoder and decoder sub-networks were connected through a series of nested, dense skip pathways which could capture features at different levels and integrate them through feature overlay. Res U-Net and Dense U-Net were inspired by residual connection and dense connection, respectively, and replaced each sub-module of U-Net with a form with residual connection and dense connection [31]–[33]. Res U-Net deepens the number of network layers and adds more skip connection between the encoder and decoder, so that it can better combine the background semantic information of the image and perform multi-scale segmentation.

Besides, inspired by the multi-scale feature extraction structure such as FPN, we find that high-level features combined with low-level features are widely used in semantic segmentation and achieve good results. On the basis of FPN, Pyramid Attention Network (FPA) uses convolution kernels to further encode low-level features before fusing low-level features and high-level features [34]. Chen et al proposed the Atrous Spatial Pyramid Pooling (ASPP) module in order to capture multi-scale information. ASPP can effectively mine convolution features of different scales and fuse them together by using hole convolution with different sampling rates [35]. Therefore, effectively fusing various information can effectively improve the detection effect of the network. Inspired by this, we creatively combine the concept "high-level features combined with low-level features" and "channel attention model", and propose our Cross-level Feature Fusion Module (CFFM).

III. PROPOSED METHOD

In this section, the proposed method is introduced in detail. As shown in FIGURE 2, the first step is data preprocessing and data augmentation. Data augmentation is used to enhance robustness of the network. The second step is to train the proposed Channel-attention U-Net. After training, the parameters of the network can learn the features of esophagus and esophageal cancer, and output the predicted segmentation results through up-sampling. Finally, the proposed 3D semi-automatic segmentation method loads the saved network and segments 3D esophagus and esophageal cancer.

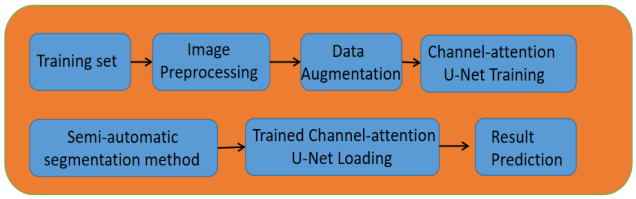


FIGURE 2. Flowchart of the proposed method for the segmentation of esophagus and esophageal cancer.

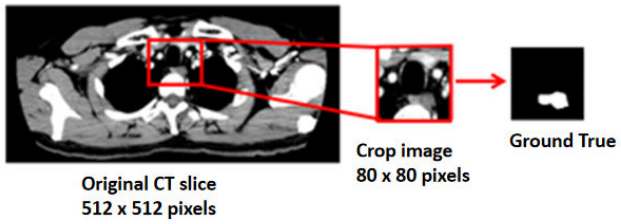


FIGURE 3. Illustration of the input data of network.

A. IMAGE PREPROCESSING

First, we convert CT images (Digital Imaging and Communications in Medicine, DICOM) into 512 × 512 bitmap format image. As shown in FIGURE 3, the size of the rectangular area containing the esophagus or esophageal cancer is 80×80. This rectangular area is used as the input to the proposed Channel-attention U-Net for each CT slice. The label is the area that esophageal or esophageal cancer area marked by the doctor, and it is also the desired output of the network.

B. DATA AUGMENTATION

In deep learning, data augmentation is conducive to enhancing the robustness of the network. Through operations such as cropping and image translation, the network can identify esophagus and esophageal cancer that appear in different positions in the rectangle. To improve the robustness of the segmentation algorithm, we transform and crop each training data and test data in multiple directions, respectively, to generate multiple rectangles. Because the esophageal or esophageal cancer area can appear anywhere in the medical image and not just the center of the selected rectangle. As a result, both the training and test sets have been augmented by almost five times as before.

C. TRAINING AND TESTING

The preprocessed esophagus and esophageal cancer images are input to the proposed Channel-attention U-Net directly,

and their corresponding label images are used as the supervision of the network.

For the details of the training phase, we use Adam as our network optimizer and the training epoch is 30 [36]. The batch size is 4. In theory, we use cross-entropy as a loss function, but consider that this is a binary classification problem with esophagus or esophageal cancer as the foreground and the rest as the background, so we use a Binary Cross-Entropy (BCE) function as the applied loss function [37]. Since the input value of the BCE function cannot be negative, we use softmax function to make the value of the probability map of the network output between 0 and 1.

The shapes of the esophagus and esophageal cancer are very different. In addition, esophageal and esophageal cancer does not have a fixed spatial relationship with the surrounding tissues from the dataset, which results in very different features between them. Therefore, some features are easy to learn by the network, and some features that are difficult to learn by the network. In other words, due to the large differences in features, the network can fit some features quickly and fit other features very slowly.

If all data are used for training in one iteration, then the network will be over-fitting in some data and under-fitting in other data. To solve this problem, we raise the threshold of the loss value which is set to 0.04 after many experiments. The proposed network performs training only when the loss value of a batch of data is greater than the threshold value.

In the testing phase, in order to verify the generalization ability of the network model, we add some similar but different images to the training set in the test set. To compare the pros and cons of some existing models with ours in generalization, we select 100 samples from the test set as the extra test sets. These samples are not in the training set, but its features are in the training set. Then we compare some state-of-the-art deep learning models with ours.

D. ARCHITECTURE OF CHANNEL-ATTENTION U-NET

In this section, the network architecture of Channel-attention U-Net is presented (illustrated in FIGURE 4). This model is based on U-Net which is widely used in medical image segmentation tasks and pixel-level classification because of its superior performance. On this basis, we embed a Cross-level Feature Fusion Module (CFFM) to form our Channel-attention U-Net. Channel-attention U-Net is an end-to-end network architecture, which consists of two main parts. The first component is a Cross-level Feature Fusion Module (CFFM), which consists of several Channel Attention Modules (CAM). The CFFM can fuse the features of the top-level feature map layer by layer into the feature map of the bottom layer, and achieve the purpose of guiding the feature selection of the bottom-level feature map by using the CAM. The second component is the U-Net which is the encoder-decoder structure that the first half is used for feature extraction, and the second half is up-sampled. Especially, there is a skip connection structure between the encoder and decoder. The skip connection links the corresponding

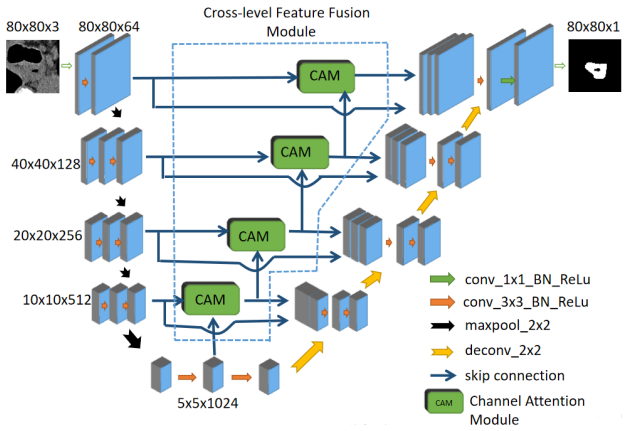


FIGURE 4. Detailed architecture of the proposed Channel-attention U-Net.

down-sampling and up-sampling feature maps. This process can solve the problem of relationship with the surrounding tissues from information loss caused by down-sampling and improves focus on convolution feature extraction.

The convolution kernel in the proposed Channel-attention U-Net network is 3×3 except the last layer is 1×1. In the encoder phase of the proposed network, the number of channels in the input image is 3. Due to the role of the convolution kernel in the encoder, the input channels increase layer by layer with the depth of the network, which are 64, 128, 512 and 1024 respectively and the output of each layer of the encoder will be used as the input of the CFFM and the input of the corresponding skip connection of the decoder. The output feature map of CFFM, the corresponding skip connection feature map, and the feature map extracted by up-sampling from the decoding layer are merged through the concatenation operation. Then perform channel dimension reduction and feature extraction through a 3×3 convolution operation.

Down-sampling is carried out by the max pool operation. Size of the pooling filter is 2×2. The purpose of down-sampling is to increase the receptive field of the feature map, so that the feature map can obtain the global features of the image and express the abstract features.

Up-sampling is carried out by deconvolution. The size of deconvolution is 3×3. The purpose of deconvolution is to expand the feature map reduced by the down-sampling operation to the original resolution layer by layer, so that feature maps from the encoder and decoder in the same stage can be concatenated together for subsequent convolution operations.

In the decoder stage, the up-sampling by deconvolution is started from the feature map with the lowest resolution. And the number of channels of the feature map is reduced from 1024 layer-by-layer to the original 64. Finally, the 1×1 convolution kernel reduces the number of channels to 1. On the one hand, a 1×1 convolution kernel can reduce the number of channels. On the other hand, the 1×1 convolution can be used to improve the robustness of the network and be used in deep networks.

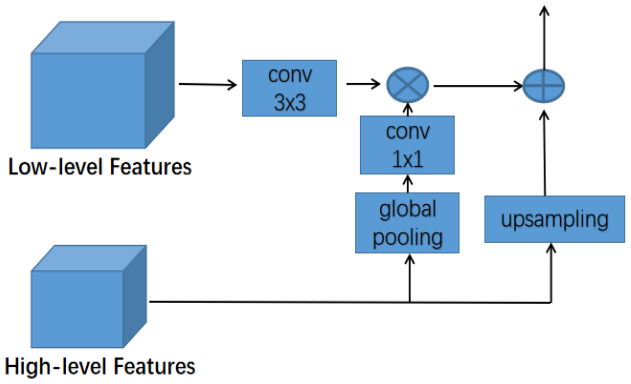


FIGURE 5. The illustration of Channel Attention Module (CAM). The high-level features guide the low-level features to emphasize or suppress related features.

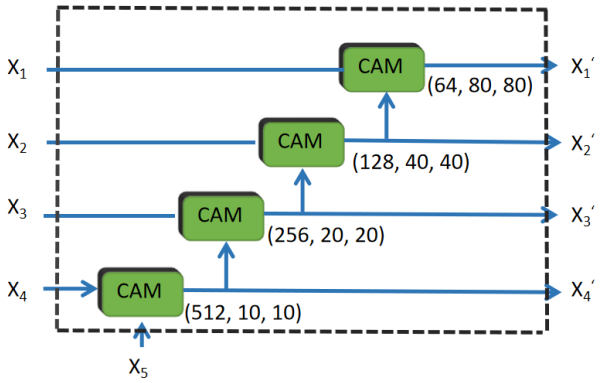


FIGURE 6. Architecture of Cross-level Feature Fusion Module (CFFM) which contains four Channel Attention Module (CAM).

1) CHANNEL ATTENTION MODULE

In this section, a Channel Attention Module (CAM) is presented, and the architecture is shown in FIGURE 5.

Since the high-level is rich in semantic information, this can help guide the selection of low-level, so as to achieve more accurate resolution information selection. Through CAM, the network can learn the weight of each channel. Thereby generating attention in the channel domain. The process of CAM is defined as:

$$X_{cam} = CAM(X_L, X_H) \tag{1}$$

where x_{cam} represented the output of CAM module. X_L and X_H represented low-level feature map and high-level feature map respectively.

$$x_{cam}^{i,j,k} = x_{gL}^{i,j,k} + \mathrm{up}(x_H^{i,j,k}) \tag{2}$$

where $x_{cam}^{i,j,k}$ is the pixel value of the ith row, jth column, and kth channel in the output feature map of the CAM. In addition, $x_H^{i,j,k}$ is the pixel value of the ith row, jth column, and kth channel of the high-level feature map. $up(\cdot)$ means bilinear interpolation.

$$X_{gL} = conv1([X_L^{'k} \times g(X_H^k)]_{k=0}^{k=c-1}) \tag{3}$$

where X_{gL} represented the feature map processed by the global pooling operation and X_L represented the low-level feature map. Because the feature map size and channel number of the low-level feature and the high-level feature are different, the low-level feature needs a 3×3 convolution operation to adjust the feature map size and merge the channel feature.

$$X_L' = conv3(X_L) \tag{4}$$

where $g(\cdot)$, $conv3(\cdot)$ and $conv1(\cdot)$ respectively represent the global pooling operation, the convolution operation with a 3×3 convolution kernel, and the convolution operation with a 1×1 convolution kernel. c denotes the total number of channels.

CAM first performs global pooling to provide a global context as a guidance of low-level features to select category localization details. Concretely, we perform 3×3 convolution filter on the low-level features to reduce channels of feature maps from lower-level feature output in U-Net encoder. The global context generated from high-level features is carried out by a 1×1 convolution operation with batch normalization and ReLU non-linearity function, then multiplied by the low-level features. Finally, high-level features are added with the weighted low-level features. This CAM module deploys different scale feature maps more effectively and uses high-level features provide guidance information to low-level feature maps.

2) CROSS-LEVEL FEATURE FUSION MODULE

Cross-level Feature Fusion Module (CFFM) consists of several Channel Attention Modules (CAM). The input of each CAM module is the output of different stages of each U-Net encoder. The output of each CAM module is not just connected to the decoder of U-Net like the skip connection, but also as a high-level feature input of the previous CAM module. The network architecture of CFFM is illustrated in FIGURE 6.

CAM can use high-level features as a guide to emphasize or suppress low-level features. We are inspired by Pyramid Attention Network (FPA) [33]. It combines the high-level features and low-level features to promote the performance on detecting small objects and focus on detailed information about objects, such as borders and textures. We take the output feature map of each encoder as the input of the corresponding CAM. Moreover, the output of each layer of CAM is used as the input of the previous CAM. Therefore, when each CAM performs channel screening, it will be guided by various higher-level features. CAM is defined as following:

$$\begin{cases} X_i' = CAM(X_{i+1}', X_i), & (i < n) \\ X_i' = CAM(X_{i+1}, X_i), & (i = n) \end{cases} \tag{5}$$

where $CAM(\cdot)$ represented the process of CAM in Equation (1). X_i and X_i' are the input and output of the CAM,

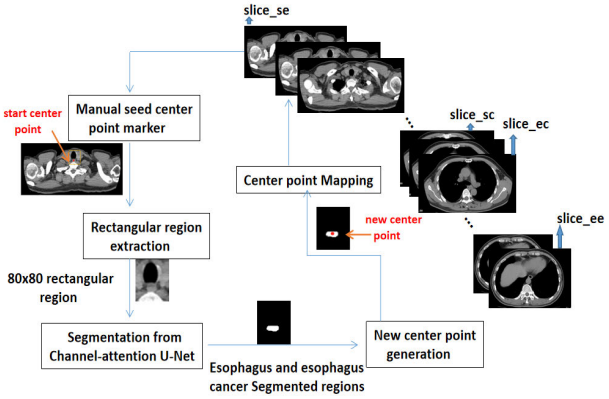


FIGURE 7. Scheme of 3D semi-automatic segmentation for esophagus and esophageal cancer: start center point is the center point of the first CT slice. slice_ce and slice_ee are the start and end CT slice of esophagus. slice_sc and slice_ee are the start and end CT slice of esophageal cancer.

respectively. Moreover, i is the current number of layers, and n denoted the number of CAM module.

E. 3D SEMI-AUTOMATIC SEGMENTATION

The esophagus is a tubular structure. Therefore, 3D esophageal cancer image processing can help doctors diagnose more effectively. As shown in FIGURE. 7, a semi-automatic segmentation method for segmenting 3D esophagus or esophageal cancer in CT images is proposed. This method is used to semi-automatically segment the esophagus and esophageal cancer in a 3D CT sequence. At the beginning of the semi-automatic segmentation method, there are five parameters that need to be set manually. They are the start and end slices of the esophagus and esophageal cancer, and the center point of the esophageal region in the first esophageal CT slice. Among them, the start center point represents the center point of the esophagus region of the first CT slice of the esophagus. It is used as the center to determine an 80×80 rectangular region as the input of the trained Channel-attention U-Net. And then, extract its center point from the segmented area and map this point to the next slice to continue to produce 80×80 rectangular area, and use it as input to the segmented network. The premise of this method is that the section thickness of the esophagus and esophageal cancer is small enough that the center of the upper esophagus is located in the area of the lower esophagus. It is suitable for most situations. The esophagus or esophageal cancer area of each CT slice is reserved for 3D rendering.

IV. EXPERIMENT

A. DATASET ANALYSIS AND EXPERIMENTAL ENVIRONMENT

All CT images used in the experiment are supported by the Department of Medical Imaging, Sun Yat-sen University Cancer Center. Chest CT scans of a total of 152 patients diagnosed with esophageal cancer were used for experimental esophageal segmentation. The parameters of each CT image are as follows: Field of view = 376mm × 376mm

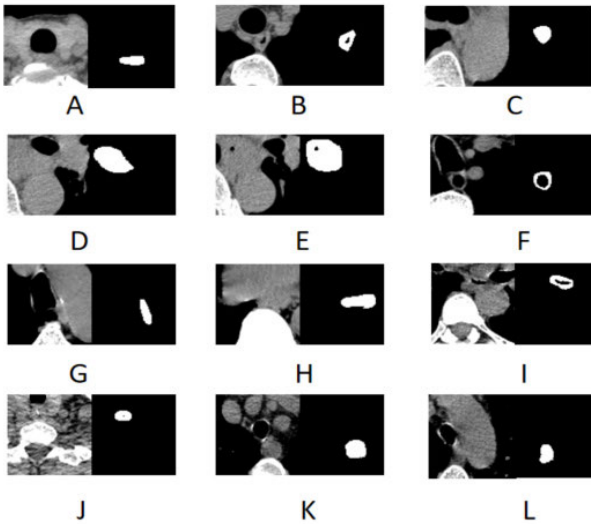


FIGURE 8. Various forms of the esophagus and esophageal cancer. the left is the input image, and right is the mask image. (A&B) esophagus under the trachea with no air hole and with air hole. (C-E, L) Large esophageal cancer besides the trachea with air hole and with no air hole. (F&G) perforated esophagus and strip esophagus around the lungs and trachea. (H&I) Esophagus not near the trachea. (J) Esophagus in different modes with others. (K) There are multiple areas similar to the esophagus.

and matrix = 512 × 512. Slice thickness = 1mm, and bits stored = 12. The CT images of esophageal cancer and esophagus were delineated by two experienced doctors as the Region of Interest (ROI). A total of 13,050 CT images were manually labeled.

We selected 40 sets of the total 45 sets of CT scans as the training set and validation set, and the rest as the test set. The CT images of the esophagus or esophageal cancer areas are delineated manually by doctors. A total of 1,450 slices in the test set are also delineated.

Data augmentation was implemented on training and test data. Each CT slice generates five rectangles containing ROI for training proposed Channel-attention U-Net. Then we divide a part from the training set at a ratio of 0.2 as the validation set. That is 46,400 samples in the training set. 11,600 samples in the validation set and 7,250 samples in the test set.

The training set contains CT images of various forms of the esophagus and esophageal cancer, and the esophagus tissue on these CT images does not have a specific positional relationship with the surrounding tissues, as shown in FIGURE 8, Since esophagus and esophageal cancer can neither be determined by judging the position of the trachea or some specific tissues, nor by the shape, which brings a huge challenge.

All the experiments were run on the following environment: UBUNTU 16.04, python 3.6 and GTX1080TI.

B. EVALUATION METRIC

The following three pixel-level measurements are used to compare the segmentation performance of the proposed approach with those of other methods: Intersection over

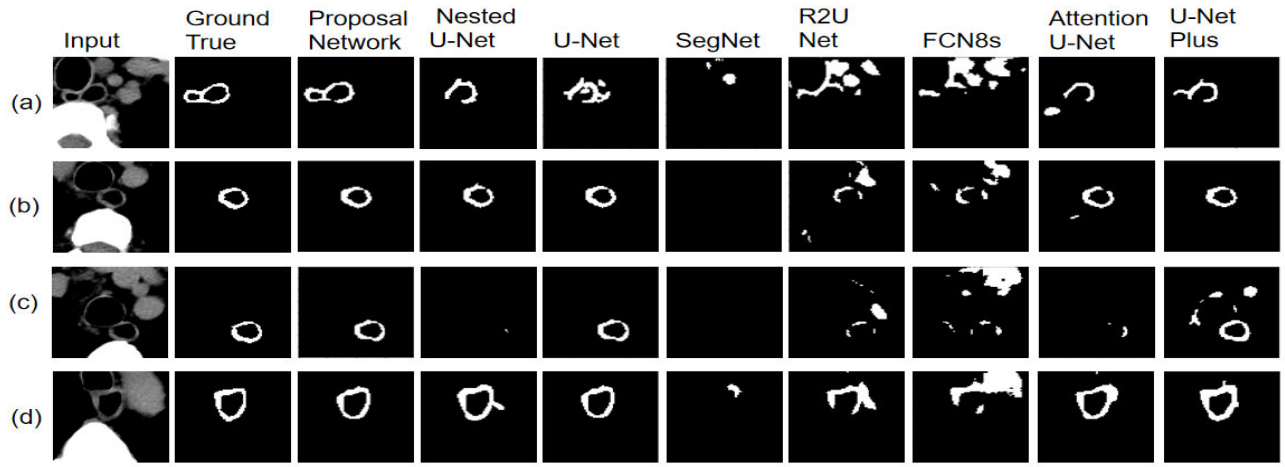


FIGURE 9. Comparison of esophagus segmentation results between the proposed model and seven other deep learning models.

Union (IoU), Dice Value (DV) and Hausdorff Distance (HD) [38].

IoU is defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{FN} + \text{FP} + \text{TP}} \tag{6}$$

where TP represents true positives, that means the overlap area of the target area manually marked by the doctors and the predicted output area of the proposed network. FN represents false negatives, that means the prediction result of this area is background, but it is actually the area of the esophagus or esophageal cancer. FP represents false positives, that means the area was erroneously segmented as esophagus or esophageal cancer by the proposed network. IoU means the ratio of the intersection and union of two sets. IoU score is a popular measure of an algorithm's pixel-level image segmentation performance. The value of IoU ranges from 0 to 1. The larger the IoU, the higher the overlap between the two regions, and the lower it is.

The second evaluation metric DV score is defined as follows:

$$DV(A_S, A_T) = \frac{2A_{TS}}{A_S + A_T} \tag{7}$$

where A_T is the mask area marked manually by doctors, and A_S is the result area segmented by model. A_{TS} represented the overlap area of A_T and A_S. DV measures the overlapping degree between target area and segmented area, and its range is in [0, 1]. The larger the DV is, the higher the consistency between mask and segmentation is.

The third evaluation metric is Hausdorff distance which is a measure describing the degree of similarity between two sets of points. Hausdorff distance (HD) is defined as:

$$HD(S, T) = \max(h(S, T), h(T, S)) \tag{8}$$

Bidirectional Hausdorff distance $HD(A, B)$ is the greater of the unidirectional distances $h(A, B)$ and $h(B, A)$. $h(S, T)$ and $h(T, S)$ are defined as:

$$\begin{aligned} h(S, T) &= \max_{s \in S}(\min_{t \in T} ||s - t||) \\ h(T, S) &= \max_{t \in T}(\min_{s \in S} ||t - s||) \end{aligned} \tag{9}$$

where $||\cdot||$ is the normal form of the distance between the point set S and T point set. And s is the point in the segmented contour (S). t is the point of the target contour (T). $h(S, T)$ actually first sorts the distance $||s - t||$ between each point s in the point set S to the point t in the T set closest to this point s, and then takes the maximum value in this distance as the $h(S, T)$ value. So is $h(T, S)$.

C. RESULTS AND ANALYSIS

The proposed model was compared with 6 state-of-the-art models including U-Net, U-Net++, U-Net Plus, SegNet, R2U-Net, Attention U-Net and FCN to achieve better segmentation of esophagus and esophageal cancer as shown in FIGURE 9 and FIGURE 10. The seven models used for experimental comparison are roughly the same in the encoder and decoder structures, but U-Net and its variants perform better than non-U-Net structures. The reason for the difference in effect is that U-Net and its variants integrate the skip connection and up-sampling feature maps in the decoder by the concatenate operation, while other networks use the adding operation. The adding operation causes the data dimension to drop and the data information to be lost. R2U-Net has more network parameters because its encoder and decoder are embedded with multiple recurrent residual convolutions. This has caused R2U-Net to be prone to overfitting when processing our dataset. Compared with Attention U-Net, U-Net++ and our network, U-Net often treats a part of the air hole as a boundary when segmenting an esophagus or esophageal cancer with an air hole. Moreover, U-Net cannot segment small air holes when segmenting the esophagus with air holes. This is due to U-Net's underutilization of detailed features. The purpose of U-Net ++ and Attention

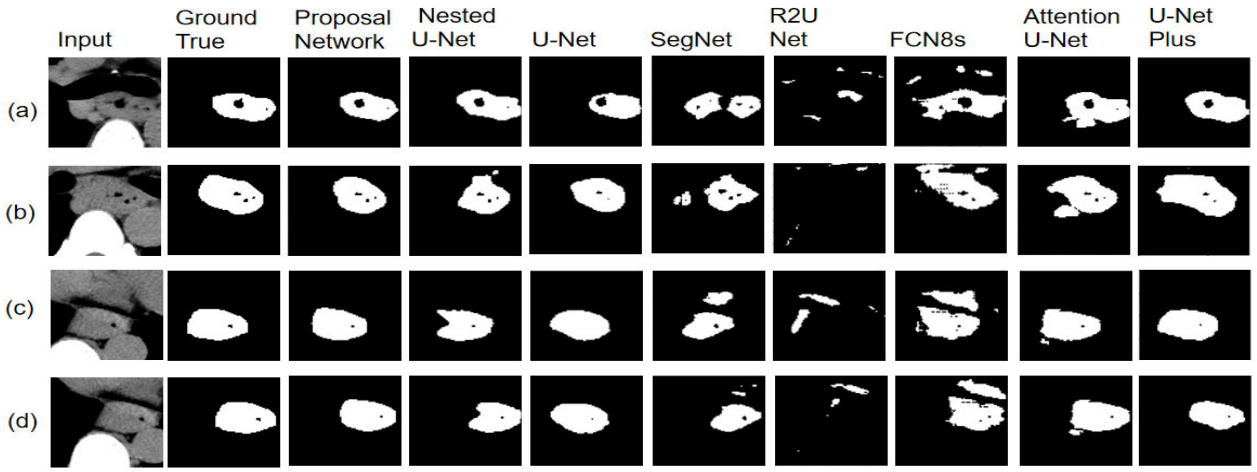


FIGURE 10. Comparison of esophageal cancer segmentation results between the proposed model and seven other deep learning models.

TABLE 1. Comparison between traditional image processing method for esophagus segmentation and the proposed method.

Methods \ Measures	IoU	DV	HD (mm)
Markov chain model [3]	0.548	0.652	3.587
Atlas selection [4]	0.537	0.640	7.301
Skeleton-shaped model [5]	0.580	0.682	7.785
Channel-attention U-Net	**0.625**	**0.739**	**3.177**

U-Net is to make full use of the feature information of each level. The former introduces features at each stage and filters them through the training process, while the latter filters features by spatial attention. However, neither of them can fuse the features of different layers with each other and guide selection of low-level features with high-level features as the proposed network. Our network not only uses Channel Attention Module (CAM) to fuse the features of different stages, but also uses Cross-level Feature Fusion Module (CFFM) to guide the selection of low-level detailed features through high-level semantic features, so that the network establishes the relationship between the spatial details of the esophagus corresponding to the shape and boundary. Therefore, the purpose of enhancing the segmentation effect and the generalization ability is achieved.

The overall effect of the experiment is good, but the problem is that our network's segmentation effect on some small boundaries is still not good enough. TABLE 1 and TABLE 2 show the average segmentation results of the proposed method compared with the existing methods in 7,250 test samples. From TABLE 2, we can see that the proposed method achieves better results than the traditional methods. From TABLE 2, our method performs better than several variants of U-Net, FCN and SegNet. Channel-attention U-Net achieves the best segmentation performance with the highest IoU (0.625), DV (0.732) and lowest HD (3.193 mm).

TABLE 2. Comparison between traditional image processing method for esophagus segmentation and the proposed method.

Methods \ Measures	IoU	DV	HD (mm)
FCN [7]	0.501	0.631	3.682
SegNet [8]	0.294	0.340	4.222
U-Net [10]	0.512	0.645	3.531
U-Net Plus [11]	0.586	0.695	3.418
Attention U-Net [12]	0.576	0.682	3.392
R2U-Net [17]	0.208	0.240	4.434
Nested U-Net [18]	0.497	0.615	3.584
Channel-attention U-Net	**0.625**	**0.732**	**3.193**

TABLE 3. Comparison of generalization capabilities between proposed model and existing state-of-the-art deep learning models.

Methods \ Measures	IoU	DV	HD (mm)
FCN [7]	0.210	0.248	5.071
SegNet [8]	0.002	0.006	5.002
U-Net [10]	0.574	0.680	4.107
U-Net Plus [11]	0.531	0.634	3.330
Attention U-Net [12]	0.399	0.483	4.085
R2U-Net [17]	0.262	0.326	4.389
Nested U-Net [18]	0.556	0.670	3.712
Channel-attention U-Net	**0.612**	**0.725**	**3.644**

The effectiveness and generalization ability of the proposed network is discussed from an additional experiment. As shown in FIGURE 11, we can see intuitively that our

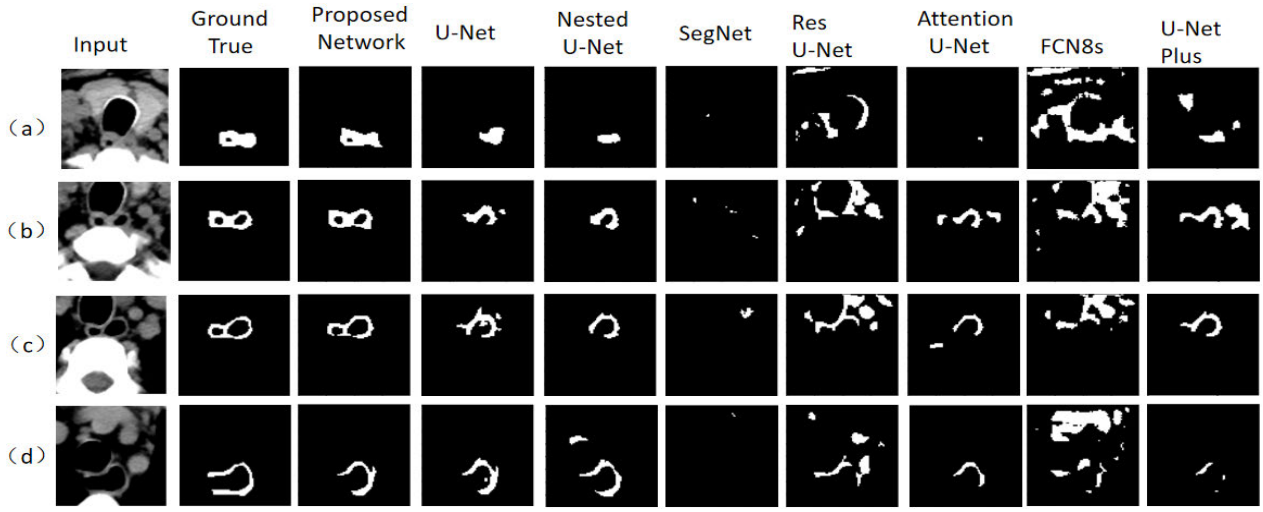


FIGURE 11. Visualization of segmentation results for esophagus. Comparative experiments on the generalization capabilities of proposed network and seven other deep learning models.

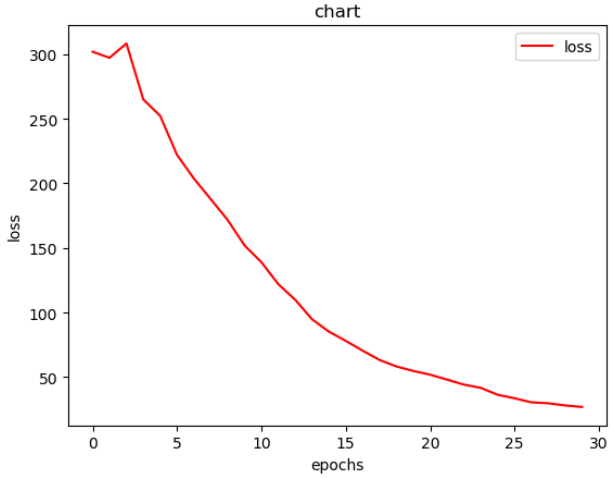


FIGURE 12. Convergence curve of our model's loss function. After 30 iterations of training, the loss value was reduced from 301 to 26.

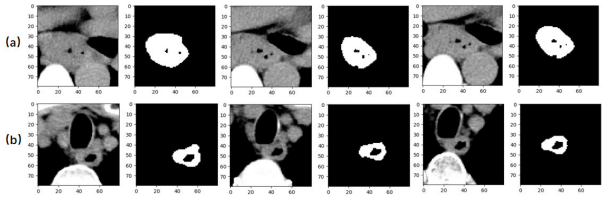


FIGURE 13. Visualization of 3D semi-supervised segmentation: (a) and (b) represent continuous esophageal cancer and normal esophagus, and their respective segmentation results.

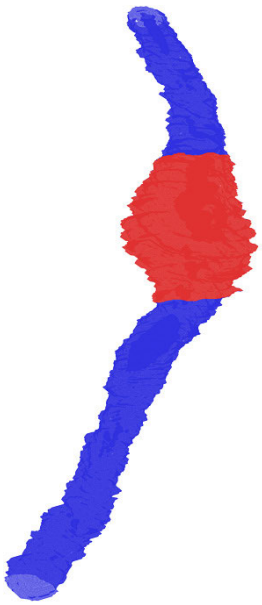


FIGURE 14. 3D rendering of the esophagus and esophagus cancer: the blue part is the normal part of the esophagus and the red part is the tumor part of the esophagus.

model performs better than the other seven models in the additional test samples.

The samples used to test the generalization of the model have similar features to the samples in the training set, but have large differences in shape. TABLE 3 shows comparisons of the generalization ability of our model with the other seven models. From the data in TABLE 3, we can see that the proposed model can obtain considerable results in the metrics of IoU (0.612) and DV (0.725). At the meantime, the results of the segmentation were verified by the doctors and proved to be consistent with the clinical diagnosis. In addition, we have invited medical experts to re-verify the 46 sets of patient data. The results prove that the diagnosis results of the proposed algorithm are all correct in 46 cases. As shown in FIGURE 12, the convergence curve of the network shows that the value of the loss function converged from 301 to 26 after 30 iterations.

3D segmentation technology has been widely used in clinical preoperative diagnostic evaluation. The application of 3D

segmentation technology in the preoperative diagnosis of esophageal cancer patients can improve the diagnosis and treatment of esophageal cancer, and provide a more accurate and scientific method for preoperative evaluation of patients. Esophageal and esophageal cancer start and end times are determined by experienced doctors. As shown in FIGURE 13, the proposed semi-automatic segmentation method proposed is used to segment continuous esophageal CT slices. As can be seen from the figure, the segmentation effect of several adjacent slice images is similar, forming a ''smooth'' gradient. This can also explain from the side that the 3D semi-automatic segmentation algorithm is robust. As shown in FIGURE 14, the segmentation results can be used for 3D rendering after stacking. The 3D model generated by the rendering operation can be rotated arbitrarily. Doctors can more clearly observe the status of esophageal or esophageal cancer and provide further assistance in the diagnosis of esophageal cancer.

V. CONCLUSION

Effective normal esophageal and esophageal cancer segmentation enables 3D rendering, which can help doctors diagnose and specify treatment options. In this paper, we have developed a deep learning architecture, named Channel-attention U-Net, which filters out the meaningful features of each level through the higher-level features and fuse the high-level features into the low-level features from top to bottom. On the one hand, after filtering features through the Channel Attention Module (CAM), the network can emphasize important features and suppress unimportant features. On the other hand, after fusing the filtered high-level features and low-level features through Cross-level Feature Fusion Module (CFFM), the network can better learn the morphological features of esophagus and esophageal cancer. Moreover, we have come to a conclusion that in terms of network generalization capabilities, models based on the U-Net structure will have stronger generalization capabilities than models with non-U-Net structure. The generalization ability of the Attention U-Net using the spatial attention mechanism was worse than that of the U-Net without the introduction of the spatial attention mechanism, which indicated that the network effect would not be better if the attention model was not introduced. Our model is better than the other seven models in the three indicators of IoU, DV and HD. The trained network will be used in a semi-automatic segmentation method to segment 3D esophagus and esophageal cancer. The segmentation results are then used for 3D rendering reconstruction to assist the doctor in diagnosis.

Through a lot of experiments, we found that the segmentation effect of the proposed network on some thin-bounded networks was not good enough. If there is a very small gap between other similar tissues and the esophagus, sometimes parts of a similar tissues are seen as prediction regions. Conversely, if there is a small boundary in the esophagus with air hole that is blocked by other tissues, the network will ignore that portion of the boundary. In the future, we will improve the edge enhancement capabilities of the network to solve this problem.

ACKNOWLEDGMENT

All experimental datasets and technical support in this paper are from Sun Yat-sen University Cancer Center. The authors thank Dr. Haojiang Li, Dr. Lizhi Liu, and Dr. Jian Zhou.

REFERENCES

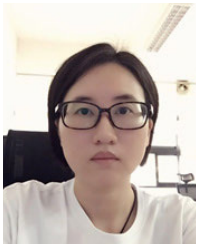
[1] B. Gupta and N. Kumar, ''Worldwide incidence, mortality and time trends for cancer of the oesophagus,'' *Eur. J. Cancer Prevention*, vol. 26, no. 2, pp. 107–118, Mar. 2017.
[2] M. Rousson, Y. Bai, C. Xu, and F. Sauer, ''Probabilistic minimal path for automated esophagus segmentation,'' in *Proc. Med. Imag., Image Process.*, vol. 6144, 2006, Art. no. 614449.
[3] J. Feulner, S. K. Zhou, A. Cavallaro, S. Seifert, J. Hornegger, and D. Comaniciu, ''Fast automatic segmentation of the esophagus from 3D CT data using a probabilistic model,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer-Verlag, 2009, pp. 255–262.
[4] J. Yang, B. Haas, R. Fang, B. M. Beadle, A. S. Garden, Z. Liao, L. Zhang, P. Balter, and L. Court, ''Atlas ranking and selection for automatic segmentation of the esophagus from CT scans,'' *Phys. Med. Biol.*, vol. 62, no. 23, pp. 9140–9158, Nov. 2017.
[5] G. Damien, C. Petitjean, B. Dubray, and S. Ruan, ''Esophagus segmentation from 3D CT Data Using Skeleton Prior-Based Graph Cut,'' in *Comput. Math. Methods Med.*, vol. 2013, pp. 1–6, Jul. 2013.
[6] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[7] Z. Hao, J. Liu, and J. Liu, ''Esophagus tumor segmentation using fully convolutional neural network and graph cut,'' in *Proc. Chin. Intell. Syst. Conf.*, 2018, pp. 413–420.
[8] A. Mittal, R. Hooda, and S. Sofat, ''LF-SegNet: A fully convolutional encoder–decoder network for segmenting lung fields from chest radiographs,'' *Wireless Pers. Commun.*, vol. 101, no. 1, pp. 511–529, Jul. 2018.
[9] B. Khagi and G.-R. Kwon, ''Pixel-label-based segmentation of cross-sectional brain MRI using simplified SegNet architecture-based CNN,'' *J. Healthcare Eng.*, vol. 2018, pp. 1–8, Oct. 2018.
[10] O. Ronneberger, P. Fischer, and T. Brox, ''U-net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[11] 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.
[12] 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
[13] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.
[14] 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.
[15] 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.*, Sep. 2017, pp. 503–506.
[16] X. Dong, Y. Lei, T. Wang, M. Thomas, L. Tang, W. J. Curran, T. Liu, and X. Yang, ''Automatic multiorgan segmentation in thorax CT images using U-net-GAN,'' *Med. Phys.*, vol. 46, no. 5, pp. 2157–2168, 2019.

[17] A. Fieselmann, S. Lautenschlger, F. Deinzer, M. John, and B. Poppe, "Esophagus segmentation by spatially-constrained shape interpolation," in *Bildverarbeitung für die Medizin*, vol. 6, no. 8. Berlin, Germany: Springer, 2008.
[18] M. Zahangir Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on U-net (R2U-net) for medical image segmentation," 2018, *arXiv:1802.06955*. [Online]. Available: http://arxiv.org/abs/1802.06955
[19] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.
[20] X. Xiao, S. Lian, Z. Luo, and S. Li, "Weighted res-UNet for high-quality retina vessel segmentation," in *Proc. 9th Int. Conf. Inf. Technol. Med. Educ. (ITME)*, Oct. 2018, pp. 327–331.
[21] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.
[22] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.
[23] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.
[24] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018.
[26] M. Zhang, X. Li, M. Xu, and Q. Li, "Image segmentation and classification for sickle cell disease using deformable U-Net," 2017, *arXiv:1710.08149*. [Online]. Available: http://arxiv.org/abs/1710.08149
[27] B. S. Lin, K. Michael, S. Kalra, and H. R. Tizhoosh, "Skin lesion segmentation: U-nets versus clustering," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.
[28] M. Salem, S. Valverde, M. Cabezas, D. Pareto, A. Oliver, J. Salvi, A. Rovira, and X. Llado, "Multiple sclerosis lesion synthesis in MRI using an encoder-decoder U-NET," *IEEE Access*, vol. 7, pp. 25171–25184, 2019.
[29] Y. Weng, T. Zhou, Y. Li, and X. Qiu, "NAS-UNet: Neural architecture search for medical image segmentation," *IEEE Access*, vol. 7, pp. 44247–44257, 2019.
[30] L. Wang, C. Xie, and N. Zeng, "RP-net: A 3D convolutional neural network for brain segmentation from magnetic resonance imaging," *IEEE Access*, vol. 7, pp. 39670–39679, 2019.
[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[32] S. Guan, A. A. Khan, S. Sikdar, and P. V. Chitnis, "Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 568–576, Feb. 2020.
[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
[34] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," 2018, *arXiv:1805.10180*. [Online]. Available: http://arxiv.org/abs/1805.10180
[35] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[37] K. Hu, Z. Zhang, X. Niu, Y. Zhang, C. Cao, F. Xiao, and X. Gao, "Retinal vessel segmentation of color fundus images using multiscale convolutional neural network with an improved cross-entropy loss function," *Neurocomputing*, vol. 309, pp. 179–191, Oct. 2018.
[38] X. Du, W. Zhang, H. Zhang, J. Chen, Y. Zhang, J. C. Warrington, G. Brahm, and S. Li, "Deep regression segmentation for cardiac bi-ventricle MR images," *IEEE Access*, vol. 6, pp. 3828–3838, 2018.

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

JUNWEN ZHU received the Bachelor of Engineering degree from Guangzhou University, in 2017. He is currently pursuing the degree with the Computer Department, Guangdong University of Technology. His research interests include computer vision, semantic segmentation, object detection, street scene segmentation, and medical image segmentation.

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.



ZHUOWEI WANG received the Ph.D. degree in computer system architecture form 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, distributed systems, and so on.

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 with the Guangdong University of Technology, a Computer Dean of the Guangdong University of Technology, a Doctoral Tutor, an Excellent Teacher of Nanyue, and a National-Level Training Target for the Thousand-Ten Thousand Project of Cross-Century Talents in Guangdong Province. His main research interests include knowledge graph, knowledge automation, and information physics fusion systems. He is also a member of the China Computer Federation. He is also the Executive Director of the Robotics Professional Committee of China Automation Association and the Vice Chairman of the Guangdong Automation Association.

LIZHI LIU received the B.Sc. degree from North Sichuan Medical University, in 1996, and the M.Sc. and D.Sc. degrees from Sun Yat-sen University, in 2008 and 2006, respectively. He was a Visiting Associate Professor with the Department of Radiology and Imaging Sciences, Emory University. He is currently a Professor and a Radiologist with the Sun Yat-sen University Cancer Center. His research interests include oncologic imaging and medical image analysis.

HAOJIANG LI is currently a Radiologist with the State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Collaborative Innovation Center for Cancer Medicine, Department of Radiology, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China. His research interests include data mining in medical imaging and statistics.

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.