Collaborative Attention Guided Multi-Scale Feature Fusion Network for Medical Image Segmentation

Zhenghua Xu, Biao Tian, Shijie Liu, Xiangtao Wang, Di Yuan, Junhua Gu, Junyang Chen, Thomas Lukasiewicz, Victor C. M. Leung, *(Life Fellow, IEEE)*

***Abstract*—Medical image segmentation is an important and complex task in clinical practices, but the widely used U-Net usually cannot achieve satisfactory performances in some clinical challenging cases. Therefore, some advanced variants of U-Net are proposed using multi-scale and attention mechanisms. Different from the existing works where multi-scale and attention are usually used independently, in this work, we integrate them together and propose a collaborative attention guided multi-scale feature fusion with enhanced convolution based U-Net (EC-CaM-UNet) model for more accurate medical image segmentation, where a novel collaborative attention guided multi-scale feature fusion (CoAG-MuSFu) module is proposed to highlight important (but small and unremarkable) multi-scale features and suppress irrelevant ones in model learning. Specifically, CoAG-MuSF uses a multi-dimensional collaborative attention (CoA) block to estimate the local and global self-attention, which is then deeply fused with the multi-scale feature maps generated by a multi-scale (MuS) block to better highlight the important multi-scale features and suppress the irrelevant ones. Furthermore, an additional supervision path and enhanced convolution blocks are used to enhance the deep model's feature learning ability in both deep and shallow features, respectively. Experimental results on three public medical image datasets show that EC-CaM-UNet greatly outperforms the state-of-the-art medical image segmentation baselines. The codes will be released after acceptance.**

***Index Terms*—Collaborative Attention, Multi-Scale Feature Fusion Network, Medical Image Segmentation**

I. Introduction

MEDICAL image segmentation is a fundamental problem and complex task in clinical computer-aided diagnosis [4], [49], [51], which aims to help medical staff treat patients conveniently. Many deep-learning-based medical image segmentation has been introduced in the literature for different target images from computerized tomography (CT) [30]. However, due to lack of sufficient feature information compared with big organ, segmentation for small organ (e.g. pancreas) in the medical image have been a challenging task [35], [38].

The fully convolutional networks (FCNs) are the first semantic segmentation model trained end-to-end and pixels-to-pixels [29]. Many FCNs-based researches have achieved higher accuracy in medical image segmentation than the traditional methods [16]. Then, U-Net [33] is proposed for medical image segmentation. Like FCNs, U-Net is composed of encoder and decoder, and skip connections is added to remedy the information loss and high-order feature fusion. Although U-Net has already achieved some great successes, its segmentation performance for small target is still unsatisfactory. Therefore, some variant of U-Net was proposed. Our review found that the improvements based on U-Net are mainly focused on two aspects. On the one hand, it is reflected in the design of adding multi-scale modules, and on the other hand, the attention mechanism is introduced. As far as multi-scale design is concerned, U-Net^{++} [27] is a representative work. The U-Net^{++} implements dense connection realize global multi-scale fetching, but dense connection introduces complex features while adding invalid information(noise) that increase the learning difficulty of the model. In addition, the large increase of parameters need pruning operation and will decrease the performance. For the study of introducing attention mechanism in U-Net, Attention U-Net [15] achieves high performance medical image segmentation. Compared with U-Net^{++}, Attention U-Net realizes the fusion of context information through convolution and weighting on skip connection to a certain extent and improves semantic effectiveness, but it can not make use of the relationship between objects or objects in the global view, which is also very important for small object segmentation in complex background. A lot of works has proved that multi-scale network [20], [21], [35], [39] and attention mechanism [7], [41], [48] can improve the segmentation effect of the model. Since the importance of multi-scale features may also be different, an interesting idea is to integrate attention mechanism with multi-scale networks to highlight the important multi-scale features and suppress the irrelevant ones, so as to enhance the model's feature learning capability, especially on those small but orangeimportant features.

This work was supported by the National Natural Science Foundation of China under the grants 62276089, 62102265 and 61906063, by the Natural Science Foundation of Hebei Province, China, under the grant F2021202064, by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ) under Grant GML-KF-22-29, by the Natural Science Foundation of Guangdong Province, China, under Grant 2022A1515011474, by the Key Research and Development Project of Hainan Province, China, under the grant ZDYF2022SHFZ015, by the China Scholarship Council, under the number 202206700006, and by the AXA Research Fund. (Corresponding authors: Zhenghua Xu, Shijie Liu, Junhua Gu; e-mail: zhenghua.xu@hebut.edu.cn, liushijiea@163.com, jhgu@hebut.edu.cn.)

Zhenghua Xu, Biao Tian, Shijie Liu, Xiangtao Wang and Di Yuan are with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin 300131, China.

Junhua Gu is with the State Key Laboratory of Reliability and Intelligence of Electrical Equipment and the School of Artificial Intelligence, Hebei University of Technology, Tianjin, China.

Junyang Chen is with the College of Computer Science and Software Engineering and the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China

Thomas Lukasiewicz is with the Institute of Logic and Computation, Vienna University of Technology, Vienna 1040, Austria, and the Department of Computer Science, University of Oxford, Oxford OX1 3QD, United Kingdom.

Victor C. M. Leung is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China

cyan There have already existed some works that use both attention and multi-scale mechanisms for image segmentations [8], [10], [14]. In [8], a MSF-ACSA model is proposed to add additive channel–spatial attention (ACSA) modules on the skip-connections of U-Net and use multi-scale modules to achieve deep supervision at the decoder. Differently, the MCDALNet model in [10] applies the multi-scale module on the skip-connections and adds a dual attention in each layer of the decoder. However, the attention modules and multi-scale modules in these two works are used independently and in series, which are thus unable to achieve the aim of using attention modules to directly highlight the important multi-scale features and suppress the irrelevant ones in model learning. Furthermore, [14] proposes a dual-branch multi-scale attention block to obtain the attention maps and multi-scale feature maps simultaneously parallel; however, the attention maps are simply added with the multi-scale feature maps in [14] without any deep integration, making the attention maps only have limited effect on the importance of multi-scale features. Therefore, the need of a work that can achieve deep integration of attention and multi-scale mechanisms is still compelling.

Therefore, we propose a novel *Collaborative Attention Guided Multi-Scale Feature Fusion* with *Enhanced Convolution* based U-Net (EC-CaM-UNet) model for more accurate medical image segmentation. cyanThe most innovative part of EC-CaM-UNet is to propose a *Collaborative Attention Guided Multi-Scale Feature Fusion (CoAG-MuSFu)* module that can deeply fuse the attention maps generated by the attention module with the multi-scale feature maps generated by the multi-scale module using a 1x1 convolution block; consequently, the attention information and multi-scale feature information can be fused deeply and comprehensively to better highlight important multi-scale features and suppress irrelevant ones in model learning and achieve better medical image segmentation performances than the existing works [8], [10], [14] (as proved in our experiments). Specifically, CoAG-MuSF is added after the last layer of encoder and before the first layer of decoder, and composed of two blocks, *Multi-Dimensional Collaborative Attention (CoA)* block and *Atrous Spatial Pyramid Pooling Multi-Scale (MuS)* block. redThe proposed collaborative attention multi-scale fusion module, CoAG-MuSFu, is different from the existing multi-scale mechanism [23], [24], which uses an attention mechanism to weight the multi-scale features to highlight key multi-scale features and suppress irrelevant ones. orangeFor the CoA block, we estimate the importance of features in both the spatial and channel dimensions. While some previous attention mechanisms focus solely on spatial attention or channel attention, CoA integrates both aspects to capture comprehensive feature dependencies. This allows our model to attend to relevant spatial regions as well as significant channel information, enhancing the overall representation learning process. The spatial attention in CoA is computed based on a self-attention approach [41], leveraging local and global correlations between regions on the feature map. This ensures accurate weight estimation by considering both local context and global context, enabling the model to better emphasize important spatial features and suppress less relevant ones. Since the dependencies between channels are estimated using the whole feature map, instead of using pixel-wised self-attention, we adopt an one-dimensional autoencoder in CoA for channel attention computation, which, comparing to self-attention, has much lower computational complexity. The resulting weight matrix is then fused by the multi-scale feature map generated by MuS block to assign the different orangemulti-scale features with different weights, which then highlights important multi-scale features, and depresses the relevant ones. For MuS block, it takes the decoder's feature as the input, then orangeextracts the multi-scale features by applying dilated convolution with different dilation rates. The output of the MuS block is to concatenate the feature maps of dilated convolution from different scales, and reduce the channels to the original number through 1×1 convolution.

Besides, to solve the problem of low effectiveness of shallow features, we add an *Enhanced Convolution* to the single convolution kernel of U-Net to enhance the feature extraction ability of shallow stage model. At the same time, the addition of asymmetric enhanced convolution also achieves robustness to flip image learning, which is beneficial to feature decoding after shallow feature fusion. Finally, to ensure that the attention module works better, we add extra supervision path.

Our work's contributions can be summarised as follows:

- We identify the limitation of the existing medical image segmentation models, and propose EC-CaM-UNet to integrate attention mechanism with multi-scale network for better small organ segmentation.
- We propose a collaborative attention guided multi-scale feature fusion (CoAG-MuSFu) module to deeply and comprehensively fuse the attention information and multi-scale feature information, so as to better highlight important multi-scale features and suppress irrelevant ones in model learning. In addition, we further introduce an additional supervision path and enhanced convolution into the segmentation to enhance the deep model's feature learning ability in both deep and shallow features, respectively.
- Experimental results on three public medical image datasets show that (i) EC-CaM-UNet greatly outperforms the state-of-the-art medical image segmentation baselines, (ii) the components of CoAG-MuSFu are all effective and essential for the model to achieve superior performances, and (iii) CoA is much better than the state-of-the-art attention mechanisms in identifying the important but less remarkable segmentation features.

The rest of this paper is organized as follows. Section II introduce the related work. Section III details the method provided by us. Section IV summarizes the experiment setup and the results. Discussion about experiments is stated in Section V, followed by the conclusion in Section VI.

II. Related Work

orangeMedical Image Segmentation. Medical image segmentation is to identify and delineate the targeted objects, e.g., organs or lesions, in clinical images. It is essential for medical image analysis to segment organs or tissues in low contrast

and high noise images [20]. The above process needs a large number of experts, but computer-aided diagnosis and treatment can realize the automatic analysis of complex medical images at a low cost. Deep learning based methods have already been widely applied in medical image segmentation tasks. FCN [29] is the first end-to-end image segmentation model using convolutional neural networks; FCN based medical image segmentation [19], [40] is mainly achieved by using convolution and pooling operations for feature learning and then applying a deconvolutional up-sampling based skip architecture for pixel-level classifications. To obtain more refined segmentation, U-Net [33] is further proposed to upgrade FCN to a structure with symmetrical contracting (downsampling) and expansive (up-sampling) paths, and skip connections are also used in U-Net to concatenate the deep and coarse features in the expansive path with the shallow and fine features in the contracting path for more accurate and detailed segmentation. U-Net is arguably the most widely adopted deep model for medical image segmentation; recent works witness the application of U-Net in various small target segmentation tasks, e.g., segmenting pancreas [15], [35], retinal vessel [34], [43], tumors [47], [52], etc. Despite achieving some successes, the performances of the existing U-Net based deep models are still unsatisfactory, especially for segmenting the small objects in medical images, so we propose the *EC-CaM-UNet* in this work.

orangeAttention Based Methods. With the success of the attention mechanism [41] in the natural language processing (NLP) task, researchers have started introducing this idea into computer vision. Some researches [31], [32] has proved that adding the attention module to the model can effectively improve performance. Hu et al. [25] propose a Squeeze and Excitation (SE) attention mechanism, which collects the information hidden in the channel-wise, fuzzes the spatial context by average pooling, and redistributes the channels' weight through a full connected neural network. CBAM attention [37] extends SE attention to utilize information in two dimensions, i.e., spatial and channel dimensions, to estimate the importance weights of features. However, the information extracted by CBAM is too sparse in terms of the feature's spatial dimension, so the non-local self-attention mechanism [36] was adopted in CoA Module to compute weights in spatial dimension. We estimate the importance of a region by the sum of its dependencies with all regions and further extend SE attention to compute channel weights by adding a one-dimensional encoder-decoder to expand the receptive field and reduce computational complexity. To show the advantage of our CoA Module, additional experiments are conducted to compare CoA with SE [25] and CBAM [37] in medical image segmentation tasks.

Similar to our work, some existing works have also integrated attention mechanisms with deep segmentation models. Attention U-Net [15] utilizes the feature map of an encoder to strengthen the feature representation in the decoder and obtains robust performance in segmenting the medical image with shallow contrast. However, compared to our work, Attention U-Net is only spatial-based and its attention maps are obtained by simple convolution and lack the capability in utilizing global information. orangeRADC-Net [2] proposes a residual attention-based convolutional neural network that achieves good results in the direction of aerial scene classification. ACG-attention [44] proposes an asymmetric cross-guided attention network for actor and action video segmentation from natural language query. CCNet [11] proposes to cross attention repeatedly considering the row attention and column attention to obtaining global information. MS2AP [3] proposes a multi-scale stacked attention pooling mechanism for remote sensing scene classification, which can improve the model's robustness and generalization performance. PCAN [13] first extracts a set of prototypes from the spatiotemporal memory, and then retrieves rich information from past frames using cross-attention. orangeRecent advances of self-attention based Transformer [42] has been demonstrated to be also effective in semantic segmentation [6]; so SegFormer [46] combines Transformer with a lightweight multi-layer perceptron (MLP) decoder to generate multi-level features. However, these methods still demand high computational requirements. DA-Net [7] is a multi-dimensional attention-based segmentation model. Like CoA computes the feature dependencies in both channel and spatial dimensions; however, DA-Net applies self-attention in both dimensions while CoA adopts a one-dimensional auto encoder for more efficient estimation of channel dependencies. orangeDifferent from these attention mechanisms, the proposed collaborative attention (CoA) mechanism is used to assign weights to multi-scale features to better highlight critical multi-scale features and suppress irrelevant features in model learning.

Multi-Scale Based Methods. After the End-to-End segmentation method appeared, some researchers pay their attention to the multi-scale module to boost model performance [23], [26], [27], [35], [50]. Spatial pyramid pooling (SPP) [24] proposes effective pooling strategies which can get the same scale output during any input scale. The different kernel size extract the multi-scale feature at the same time. Also, using pyramid pooling, the PSP-Net exploits the capability of global context information by different-region-based context aggregation through. To detect the target with different scales, Feature Pyramid Network (FPN) [28] was proposed. Unlike the above structure, the FPN will choose the optimal one among multi-scale paths. Based on the above theory and practice, some multi-scale structures are designed to be applied to image segmentation.

orangeHR-Net [45] simultaneously connects convolution streams from high-resolution to low-resolution, while maintaining high-resolution representations throughout the process. redMSU-Net [18] employs multiple convolutional sequences to extract more semantic features from images and uses convolutional kernels with different receptive fields to make the features more diversified. DeepLabV3+ [22] proposes the atrous spatial pyramid pooling multi-scale (MuS) module, which replaced pooing with three 3×3 convolutions with different dilation rates and a 1×1 convolution, then concatenated the results with the input feature maps. A diverse atrous rate can expand the receptive field and significantly

decrease the number of parameters. redIn the field of medical image segmentation, CE-net [9] utilizes multi-scale dilated convolution to extract rich feature representations, followed by multi-scale pooling operations to extract more contextual information. Unlike MuS, a local multi-scale model, U-Net^{++} is designed with a globally densely connected multi-scale structure and supervises each scale to realize the multi-scale feature selection. The above works process the multi-scale pooling or convolution operations in parallel, then concatenate every scale to the downstream task. However, not every scale feature is valid, so stitching multiple multi-scale features together is inefficient.

cyan**Methods Using Both Attention and Multi-Scale mechanisms.** There have already existed some works that use both attention and multi-scale mechanisms for image segmentations [8], [10], [14]. Specifically, [8] proposes a MSF-ACSA model to add additive channel–spatial attention (ACSA) modules on the skip-connections of U-Net and use multi-scale modules to achieve deep supervision at the decoder of U-Net Differently, the MCDALNet model in [10] applies MCL modules on the skip-connections of U-Net to extract multi-scale contextual information with adaptive receptive fields, and add dual attention in each layer of the decoder. However, the attention modules and multi-scale modules in these two works are used independently and in series, which are thus unable to achieve the aim of using attention modules to directly highlight the important multi-scale features and suppress the irrelevant ones in model learning. Similar to our work, [14] proposes a dual-branch multi-scale attention block to obtain the attention maps and multi-scale feature maps in parallel; however, the attention maps are simply added with the multi-scale feature maps in [14] without any deep integration, making the attention maps only have limited effect on the importance of multi-scale features. Different from these existing works, the proposed CoAG-MuSFu module uses a 1x1 convolution block to deeply fuse the attention maps generated by the collaborative attention module with the multi-scale feature maps generated by the multi-scale module, which thus better highlights important multi-scale features and suppresses irrelevant ones in model learning and achieves better medical image segmentation performances than the existing works [8], [10], [14] (as proved in our experiments).

III. Methodology

Complex scenes in medical imaging, especially in abdominal images, clustered many organs with similar density. The effects of background noise such as bowel and septum make organ segmentation difficult. As mentioned above, existing neural networks can obtain local correlations but cannot capture the relationships of long-range features; some multi-scale fusion schemes can improve model performance, but not all features at all scales are practical. Therefore, we propose the EC-CaM-UNet network, which strives to deeply and comprehensively fuse attention information and multi-scale feature information, so as to better highlight important multi-scale features and suppress irrelevant features.

redOur method's overall architecture is shown in Fig. 1, where an overall diagram of the proposed EC-CaM-UNet is shown in the middle area of Fig. 1, and the detailed diagram of the new blocks, EC, CoA, MuS, and Side Output, are respectively shown in the SubFig. 1 (a)-(e). In the Enhanced Convolution Based Encoder, the inputs are first downsampled four times, where the square convolution kernel 3×3 was replaced by the 1×3, 3×1, and 3×3 kernel (The enhanced convolution is shown in Fig. 1 (a)) arranged in parallel. The asymmetrical convolution effectively eliminates the imbalance distribution of features learned by the square convolution kernel [5], which will enhance the square convolution's learning ability. orangeIn medical imaging, the complex structures and details of images are of great importance, and dilated convolution can better capture this information. Then, the dense feature maps will be fed into the CoAG-MuSFu Module, which is combined with CoA Block and MuS Block, respectively. The CoA Block evaluates feature dependencies in two dimensions in different ways. The MuS Block extracts multi-scale feature representation. Then the dependencies and multi-scale features are fused by 1×1 convolution, which can assign weights to different scale information. Finally, the CoAG-MuSFu's output as the input of Decoder with Side Output Module. orangeIn medical image segmentation, the morphological and textural differences among different tissues and organs are significant. Therefore, the utilization and proper allocation of multiscale information in feature extraction and fusion become particularly important. The design of the CoAG-MuSFu module aims to address this issue by evaluating the feature dependency relationships in different dimensions and extracting multiscale feature representations. It can comprehensively utilize feature information and allocate appropriate weights for each scale by fusing different scale information, thus improving the recognition and segmentation ability of various tissues and organs in medical image segmentation tasks.

The depth of Encoder and Decoder is five, each layer consists of two basic convolution layers (convolution layer, batch normalization layer, and ReLU layer) cascaded; the skip-connection structure between Encoder and Decoder remains the same as U-Net. All the convolution operations were followed by batch normalization, the phenomenon of gradient dispersion will be improved [12], and the convergence speed of the model will be accelerated.

A. Collaborative Attention Guided Multi-Scale Feature Fusion

The Collaborative Attention Guided Multi-Scale Feature Fusion (CoAG-MuSFu) module can deeply and comprehensively fuse the attention information and multi-scale feature information, which consists of two parts; one is the Multi-Dimensional Collaborative Attention Block, and the other is the Multi-Scale Fusion Block. Now we will introduce these two parts respectively.

1) Multi-Dimensional Collaborative Attention Block: We will introduce the Multi-Dimensional Collaborative Attention Block to model the long-range relationship in spatial and get the adaptive importance among channels simultaneously. After that, the two kinds of feature representations would be utilized to form the final "attention" feature map. As shown in Fig. 1(b), we design the Collaborative Attention

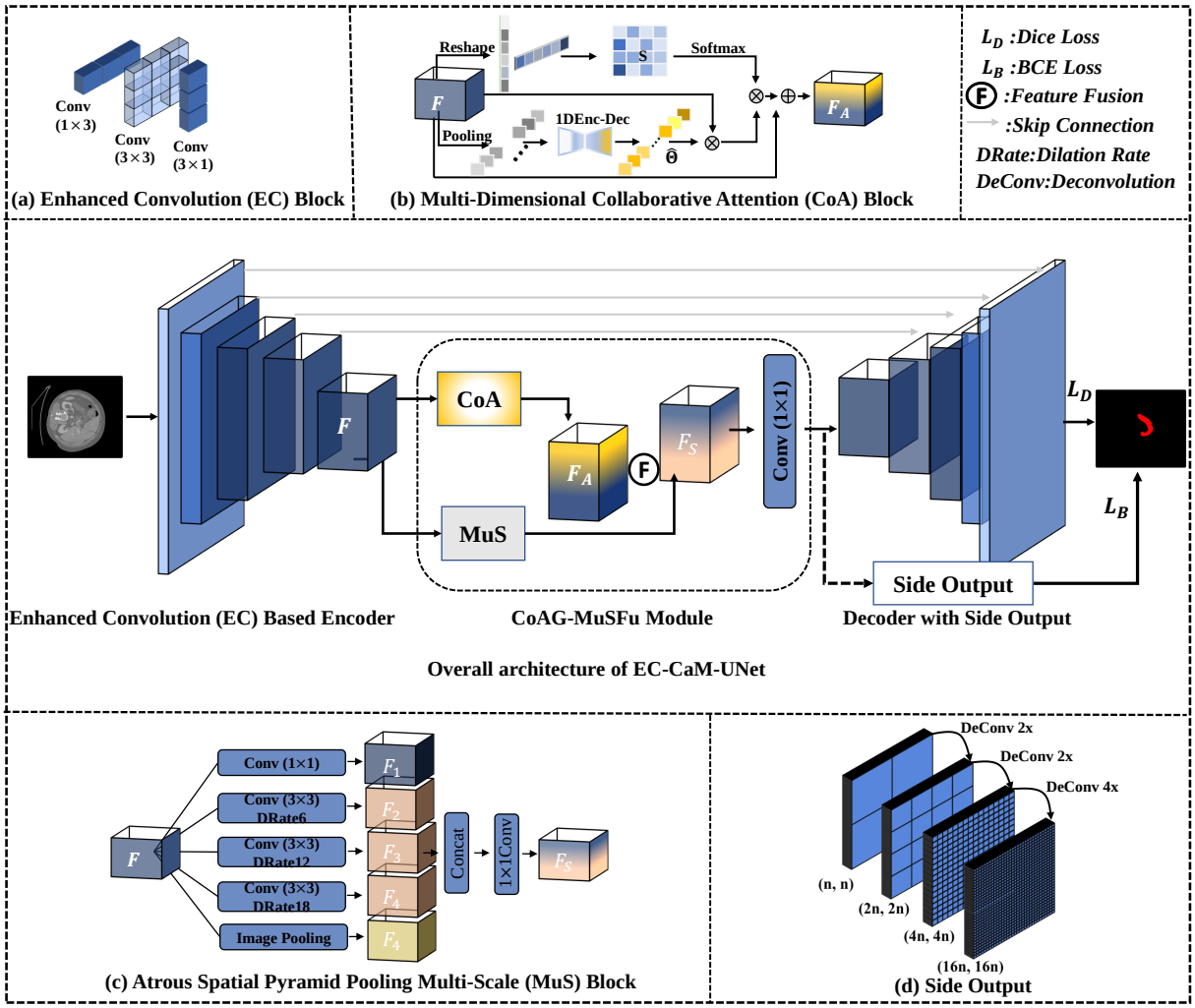


Fig. 1: redOverview and details of the proposed EC-CaM-UNet, where the overall architecture of EC-CaM-UNet is shown in the middle, the detailed diagram of Enhanced Convolution (EC) block used in the encoder of EC-CaM-UNet is shown in (a), the detailed diagrams of Collaborative Attention (CoA) block (colored by yellow) and Multi-Scale (MuS) block (colored by grey) used in the CoAG-MuSFu Module are shown in (b) and (c), and that of side output operation used in the decoder EC-CaM-UNet is shown in (d).

Block that combines of channel-correlation attention block and spatial-correlation attention block. The high dimension semantic feature from the encoder path of U-Net will be put into the Collaborative Attention Block. orangeBy utilizing the structure of the collaborative attention block, we are able to obtain more accurate feature representations, which in turn improves the ability to recognize target structures in medical images. Additionally, our method can reduce errors and noise during the segmentation process, and improve the accuracy and stability of segmentation. cyanThe CoA Block is carried out in two-step, it collects the long-range dependencies by adding spatial-correlation attention block and generating an attention map denoted as $\mathbf{S}$ in $\in R^{HW\times HW}$:

$$orange\mathbf{S} = f_{\mathcal{AS}}[\mathbf{F};\eta], \tag{1}$$

cyanwhere $f_{\mathcal{AS}}$ represents the *Channel-correlation Attention Block*, and η represents net parameters.

cyanSimultaneously, the channel-correlation attention block constrained the channels information parallel:

$$orange\hat{\Theta} = f_{\mathcal{AC}}[\mathbf{F};\mu], \tag{2}$$

$$orange\hat{\mathbf{F}_k^C} = \hat{\mathbf{F}}_k \cdot \hat{\Theta}, \tag{3}$$

cyanwhere $f_{\mathcal{AC}}$ represents the *Spatial-correlation Attention Block*, μ represents net parameters. The Θ is the vector comes from the channel-correlation attention block and $\mathbf{F}_k$ is the vector of $\mathbf{F}$ along the channel direction.

cyanThen, the feature maps $\mathbf{F}^C$ weighted by channel-correlation attention block were multiplied with matrix $\mathbf{S}$. There are better original spatial feature representations distributed in raw feature map $\mathbf{F}$. We obtain both dimensions' attention information from feature map $\mathbf{F}$, ensuring no impact between the two operations. Finally, the weighted feature map in the channel dimension will be affected by the spatial attention map to achieve the purpose of fusion.

$$orange\mathbf{F}_A = \mathbf{S} \odot \mathbf{F}^C + \mathbf{F}, \tag{4}$$

cyanwhere $\mathbf{F}^C$ is the channel-correlation attention block's output, $\mathbf{S}$ is the spatial-correlation attention map, $\odot$ represents the matrix multiplication.

In summary, we obtained channel-correlation information and spatial information from the same input feature and then realized the application of spatial information based on channel weighting, the application of residual structure made up for the lost global context. This is particularly important for medical image segmentation tasks, as many anatomical structures in medical images have similar shapes and appearance features, requiring sensitive identification of subtle feature differences.

The network with an encoder-decoder structure will take CT or MR slices as input and extract features by 2D convolution. In the encoder path, the number of channels increases with the spatial resolution decreases. Each channel of the feature map can represent one or several kinds of features and much semantic information. According to the principle of 2D convolution, the convolutional kernel's update does not directly relate between different channels when the network propagates forward, which is decided by back-propagation mainly. Therefore, we build the channel-correlation attention block to add a new connection between different channels. Adding the channel-correlation attention block will give the different weights for different channels according to the different importance and combine the relatively independent channel. We can more accurately capture the features in medical images, such as the shape and size of tumor regions. Compared to traditional methods, our approach has better performance and accuracy in medical image segmentation tasks, especially in more complex scenarios where the lesion areas are small or the background noise is high. orangeBy adding channel-wise correlation attention blocks, we can more accurately capture the features in medical images, such as the shape and size of tumor regions. Compared to traditional methods, our approach has better performance and accuracy in medical image segmentation tasks, especially in more complex scenarios where the lesion areas are small or the background noise is high. The structure of our channel-correlation attention block is displayed in Fig. 2(b). The block contains pooling, multiplication, and convolution operation. It takes multi-channel feature maps as input and output feature maps whose shape is the same as the input. First, we apply the average pooling on the input feature maps orange$\mathbf{F}$ with the shape of C×H×W in the channel dimension. Then, the pooling operation got a vector orange$\theta \in R^{C\times 1}$ and fed it into the 1D convolution operation. Every convolution is followed by batch normalization and Relu, where the 1×3 and 1×7 kernel was selected to mine the correlation between orangechannels from the θ(as shown in Fig. 2(b)).

On the one hand, 1×3 convolutions with *stride=1* utilized the adjacent channels' features; on the other hand, 1×7 convolutions with *stride=2* achieved the feature compression while expanding the receptive field further. This downsample operation with different kernel sizes was able to obtain the channel's information from its neighbors and from afar during the window sliding. Downsampling gathered the most critical information; then, the deconvolution (interpolation) would reassign it. We use Sigmoid as the last convolution layer's activation function, mapping the values between 0 and 1. Compared with the fully connected layer applied in the SENet, the convolution squeezes the number of parameters and models the relationship of the adjacent channel better. Finally, assigning the corresponding weight to each channel. The above process can be expressed as the following formula:

$$orange\theta_k = orange\frac{1}{H\times W}\cdot\sum_{i=1}^{H}\sum_{j=1}^{W}\mathbf{F}_{i,j,k}, \tag{5}$$

$$orange\hat{\Theta} = orangeConv(\{\theta_1,\theta_2,...,\theta_{k-1},\theta_k\}), \tag{6}$$

$$orange\mathbf{F}'_{i,j,k} = orange\Theta_k\cdot\mathbf{F}_{i,j,k}, \tag{7}$$

where Θ_k is the k^{th} channel's weight, and $orange\mathbf{F}_{i,j,k}$ is a element in the feature map orange$\mathbf{F}$.

As mentioned above, making the low contrast CT slices as raw inputs brings the effect that the model will make some mistakes while classifying the organ that is surrounded by some similar density tissue. Under the medical images' complex background, it is crucial that comprehend the feature representation accurately. Deep in the model, the high dimension semantic information is plentiful. The original convolution operation is limited at the size of the slice window because the too large kernel will give a great memory pressure to the GPU. Although pooling operation or bigger kernel size can expand the receptive field, it is not easy that transfer information at a distance in the feature map with convolution. To aggregate the long-range dependency and gather information, we use the self-attention method [41]. The spatial-attention block based on the self-attention mechanism is shown in Fig. 2(a). We get the relationship of one pixel with others in feature volume by calculating similarity.

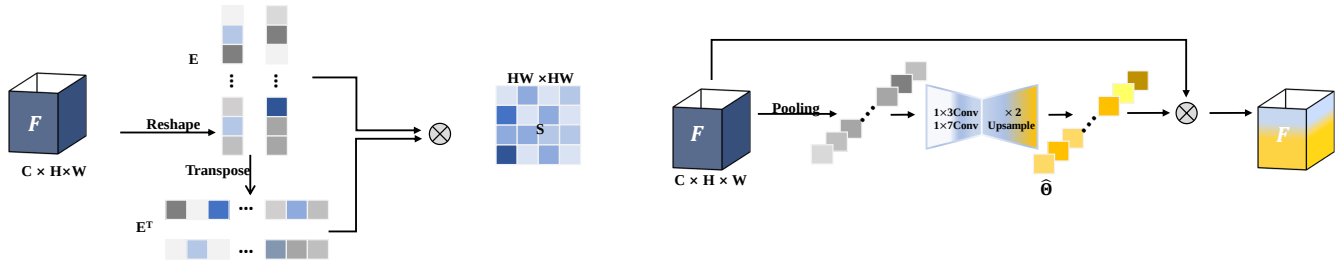
Also, taking orange$\mathbf{F}$ as input, we flattened the feature map for each channel into a vector and concatenated these vectors to get map orange$\mathbf{E} \in R^{C\times HW}$. Then, the transposition of orange$\mathbf{E}$ and orange$\mathbf{E}$ is multiplied to obtain attention map orange$\mathbf{S}$. In the process of spatial attention, we regard each channel vector as the feature expression of space points. Intuitively, we perform dimensional compression through the channel vector dot product to obtain the spatial relationship of each channel vector to each other. In the last, apply a softmax layer to the column of orange$\mathbf{S}$. The operation can formula as follows,

$$orange\mathbf{E} = orangereshape(\mathbf{F}), \tag{8}$$

$$orange\mathbf{S}_{ij} = orange\frac{\exp(\mathbf{E}^T\odot\mathbf{E})_{ij}}{\sum_{j=1}^{HW}\exp(\mathbf{E}^T\odot\mathbf{E})_{ij}}, \tag{9}$$

where $orange\mathbf{S}\in R^{HW\times HW}$, and $\odot$ is matrix multiplication.

2) Multi-Scale Module: Atrous convolution is a powerful tool that allows us to explicitly control the resolution of features calculated by the depth convolution neural network and adjust the field of view of the filter to capture multi-scale information. The introduction of the dilated convolution to control the field of view has shown promising; it can extract multi-scale contextual details. Then, Chen et al. [21] proposed



(a) Spatial-Correlation Attention Block

(b) Channel-Correlation Attention Block

Fig. 2: (a) and (b) are the details of *Spatial-correlation Attention Block* and *Channel-correlation Attention Block* respectively.

MuS Fig. 1(c), which is a parallel atrous convolution block to capture multiple-scale information simultaneously. MuS captures the contextual information at different scales, and multiple parallel atrous convolutions with varying rates in the input feature map are fused. In our network, we make the MuS module a bridge between the encoder and the decoder sections, which will perform multi-scale fusion without reducing the resolution. redThe proposed collaborative attention multi-scale fusion module, CoAG-MuSFu, uses attention mechanism to weight multi-scale features to highlight key features and suppress irrelevant features. Therefore, we use MuS to refine the useful multi-scale information for the semantic segmentation task. The process is shown in Fig. 1(c). The feature map **orangeE** extracted by the U-Net's encoder is fed into the MuS module; in the MuS module we follow the default setting, setting dilate rate as 6, 12, 18. After the dilated convolution with different dilatation rates, the obtained feature maps are spliced together, and then the number of channels is restored by 1×1 convolution and fused with the output CoA module.

3) Deep Feature Fusion: The attention information generated by the CoA Block is deeply fused with the multi-scale features generated by the MuS Block using a 1x1 convolution module. The information at different scales is thus deeply fused using 1x1 convolution with the constraint of attention information, which thus better highlights the important multi-scale features and suppresses the irrelevant ones. Formally,

$$orange\mathbf{F} = \text{Conv}_{1\times1}(\text{Concat}(\mathbf{F_A}, \mathbf{F_S})), \tag{10}$$

orangewhere $\mathbf{F}$ is the feature after deep feature fusion, $\mathbf{F_A}$ is the attention map obtained by the CoA block, and $\mathbf{F_S}$ is the multi-scale features generated by the Mus block. orangeBy deeply fusing the attention maps generated by the CoA block with the multi-scale feature maps generated by the MuS block, the proposed CoAG-MuSFu module can achieve more accurate medical image segmentation than the existing works that also use both attention and multi-scale mechanisms in clinical practices.

B. Decoder with Side Output

From the above, we would find that applying orangethe MuS and CoA modules increased the number of model parameters. More parameters need to be learned with the increase in complexity, making the model's learning more difficult. Generally speaking, it is unreasonable to complete the whole model's learning by giving it to a single loss function, and our experiment confirmed it. A common feature in the almost abdominal CT image; is that the target is small, and the background is extensive. However, the information exposed to the model by the label was the value of one or zero, and a more extensive background led to the imbalance of categories. Thus, we selected Dice-coefficient loss ($\mathcal{L}_{\text{Dice}}$) as our model's primary loss function, which helps the model capture the tiny object better. From the Table IV, we can find that Dice Loss does not always work best, and the decrease of loss value was not smooth while training the model. Nevertheless, the research showed that dice loss might lead to gradient problems in back propagation and make the training process unstable. Therefore, we use Binary Cross-Entropy loss ($\mathcal{L}_{\text{BCE}}$) as a side output path to ensure the correct learning of the Attention module while providing a relatively stable gradient for the model's learning. Formally,

$$orange\mathcal{L}_{\text{Dice}} = 1 - \frac{2\widehat{y}y}{\widehat{y}+y}, \tag{11}$$

$$orange\mathcal{L}_{\text{BCE}} = -y\log\widehat{y} - (1-y)\log(1-\widehat{y}), \tag{12}$$

$$orange\mathcal{L}_{\text{Hybrid}} = \mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{BCE}}, \tag{13}$$

where $\widehat{y}$ is the model's prediction, and y is the corresponding label. cyanHere, the two losses are combined with the same weights because our experiments prove that using weighted sum will not have significant improvements but will bring expensive additional time-cost for tuning the hyper-parameter. However, the future users can still change this equation to the weighted form if very accurate segmentation performances are needed.

IV. Experiment

A. Dataset

MSD (Medical Segmentation Decathlon Challenge) provides data on ten kinds of human organs, including the pancreas data set. There are 281 cases of CT volume and the corresponding label marked manually by specialists. The spatial resolution of each case in this data set is 512 × 512, and the number of slices is different from 37 to 750. 236 patients are divided from the total data for the training set, 40

cases for testing, and 5 cases for validation. We use the slice and pancreas annotation to train and evaluate our method.

NIHC-TCIA (National Institutes of Health Center-The C-ancer Imaging Archive) obtains 82 abdominal contrast-enhanced 3D CT scans (70 seconds after intravenous contrast injection in portal-venous) from 53 male and 27 female subjects. The CT scans have resolutions of 512x512 pixels with varying pixel sizes and slice thickness between 1.5-2.5 mm, acquired on Philips and Siemens MDCT scanners. A medical student manually performed slice-by-slice segmentation of the pancreas as ground-truth. We randomly selected 57 cases from 82 cases as the training set data, 5 cases as a verification set, and the remaining 20 cases as a test set to evaluate the model's performance.

KiTS19 is a kidney dataset that contains 210 cases' image and segmentation labels. The manual segmentation labels were provided by medical students. The image, as well as labels in this data set, were NIFT format with shape (the number of slice, height, width), where the direction of the slice is an axial view and the slice thicknesses range from 1mm to 5mm. 170 cases, 10 cases, 30 cases were split at random as the train set, validation set, and test set. We use the slice and kidney annotation to train and evaluate our method.

orange**Cardiac** [1], [17] is a publicly available medical imaging dataset that consists of MRI scans from 20 cardiac patients, includes a total of 1,350 slices from 20 different cases, with each slice sized at 320×320 pixels. All images have been annotated by medical professionals, making them a valuable resource for evaluating the performance and accuracy of automatic segmentation algorithms. Preprocessing and enhancement techniques are applied to the images to highlight the cardiac structures effectively.

B. Baselines

We compared the MG-CoA module with five sota methods, including (i) the original U-Net for medical image segmentation, (ii) a method using Attention Gate module for pancreas segmentation and that is the state-of-the-art method, (iii) a dual attention module that proposed for natural image segmentation, FCN combined with which achieved state-of-the-art on CityScape and VOC datasets, (iv) the original channel attention and spatial attention module, (v) densely connected U-Net.

1) U-Net. We can see from the Section I, the U-Net is a very important structure in medical image segmentation. Unlike the FCN method, the U-Net utilized the denser long-skip-connection and feature concatenate, which made the high and low-level semantic features integrated fully. Like most researchers studied the medical image segmentation, we selected U-Net as the baseline method.
2) Attention U-Net. Compare with the U-Net, [15] proposed the concept of attention gate and combined with the U-Net, which realized state-of-the-art in pancreas segmentation task. The attention gate module was embedded in each decoder path level, which made current level feature maps as "signal" and one level lower feature as "gate" information. Next, the "signal" and "gate" were fed into the attention gate module to format a weighted map. In the last, multiplying "signal" with a weighted map and continue to the next level.
3) DA-Net. It uses the point-wise and channel-wise attention parallel, where the idea of self-attention is used. Different from our attention module, DA module weighted the feature map in the dimension of the channel and spatial respectively and then merged the feature map.
4) DeepLabV3+. The multi-scale module is proposed in the series of deeplab. We use the MuS module as a bridge between the encoder and decoder to achieve multi-scale feature complementation. MuS is derived from the deeplab series, which is why we choose it as the baseline method.
5) UNet++. It is the orangestate-of-the-art method in the field of medical image segmentation, The dense connections enable the fusion of different depth networks and different scale features, which is close to the idea of our work.
6) SE [25] and CBAM [37] Module. The principles of the two methods are the same. Both of them applied the pooling operation to catch the global information and assigned by multiplication element wise.
7) orangeHR-Net [45] is a network architecture that simultaneously connects convolutional streams from high-resolution to low-resolution while maintaining high-resolution representations throughout the process.
8) orangeSegFormer [46] unifies Transformer with a lightweight multilayer perceptron (MLP) decoders and is a semantic segmentation framework that utilizes self-attention.
9) redCE-Net [9] uses a contextual encoder network to capture more high-level information and preserve spatial information for 2D medical image segmentation.
10) redMSU-Net [18] is a multi-scale U-Net for medical image segmentation, addressing the limitations of fixed receptive field and unknown optimal network width of the convolution kernels in U-Net.

C. orangeImplementation Details

In the experiment, a grayscale image with a single channel was fed into our network, and the image size is 512×512. When only a few training samples are available, data argumentation is the key to imparting the required invariance and robustness to the network. The probability of a random horizontal flip was set to 0.5. Randomly rotate images with rotate factor in [-10, 10]. Simultaneously, random gaussian blur is applied to extend data; fifty percent of each image is likely to be processed by Gaussian filtering. It is worth noting that the labeled data is processed simultaneously with the training data. The batch size was set to 4. We optimized our model with Adam optimizer. The initial learning rate was set to 0.00001, and the cosine annealing is used to adjust the learning rate. Weight decay was set to 0.0001 on the MSD dataset and KiTS dataset and 0.0002 on the TCIA and Cardiac dataset. We train the models 150 epochs and select the parameters with the best results on the validation

TABLE I: The performances of our method and the state-of-the-art baselines on four medical image segmentation datasets.

Method	DATASET											
	TCIA			MSD			KiTS			orangeCardiac		
	DSC	Sensitivity	Precision	DSC	Sensitivity	Precision	DSC	Sensitivity	Precision	DSC	Sensitivity	Precision
U-Net [33]	67.80	64.92	81.44	65.35	70.74	77.56	93.83	96.20	94.60	74.41	75.86	78.03
U-Net++ [25]	71.47	71.44	80.48	68.75	70.42	78.71	94.10	96.25	94.11	75.58	75.93	80.65
DeepLabV3+ [37]	68.08	62.86	80.74	69.02	71.21	73.13	94.54	93.81	96.01	78.26	77.05	79.50
Attention U-Net [15]	71.37	71.30	81.17	70.01	72.94	78.98	94.45	96.18	94.73	74.61	77.48	79.47
DA-Net [7]	69.17	67.70	80.97	70.77	72.70	78.37	94.57	96.10	94.79	77.82	73.45	**82.75**
orangeHR-Net [45]	orange71.13	orange68.43	orange80.12	orange71.03	orange73.51	orange76.38	orange94.66	orange95.53	orange96.72	orange78.43	orange77.55	orange79.33
orangeSegFormer [46]	orange69.17	orange67.70	orange80.97	orange70.77	orange72.70	orange78.37	orange94.57	orange96.10	orange95.30	orange77.27	orange72.88	orange82.23
redCE-Net [9]	red71.75	red69.74	red79.29	red69.22	red70.29	red77.97	red94.08	red94.02	**red96.78**	red78.19	red78.64	red82.13
redMSU-Net [18]	red70.34	red69.38	red79.72	red70.21	red74.50	red76.50	red94.71	red96.25	red95.34	red75.62	red75.18	red80.75
Our	**72.64**	**72.07**	**82.34**	**71.68**	**74.96**	**79.06**	**95.02**	**96.28**	95.29	**79.32**	**80.31**	82.51

set as the final model. We use the PyTorch tools to build our model and trained with two GTX2080Ti GPUs; thus, the Synchronized batch normalization is used. orangeIn our experiments, we do not use any pre-training models, i.e., all models are trained from scratch using the same initialization method and hyperparameters to ensure a fair comparison.

D. Evaluation Metrics

We used Dice-coefficient, precision and sensitivity as the evaluation indicators. The Dice is used to evaluate the overlap ratio between the predicted target and the ground truth, which is the most important index of medical image segmentation. The sensitivity shows that how many positive examples in the sample are predicted correctly. We evaluate the performance of the model on CT slice.

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{15}$$

$$DSC = \frac{2 * TP}{2TP + FP + FN} \tag{16}$$

where TP is the number of positive pixels that are correctly classified in the segmentation results; FP, false positive, is the number of negative pixels that are incorrectly classified as positive pixels; FN, false negative, is the number of positive pixels that are incorrectly classified as negative pixels.

E. orangeMain Results

According to the experiment results, we made a quantitative analysis. As shown in Table I, we can see that our proposed method proposed performed best on the four datasets. Compared to U-Net, our method showed an average performance improvement of 5.58% on both pancreas datasets (i.e., TCIA and MSD datasets). Compared with Attention U-Net, the DSC of our method is improved by about 1.28% on both pancreas datasets. Meanwhile, compared with the DA-Net, our method achieved a 2% performance improvement on TCIA and MSD datasets. orangeCompared with SegFormer, our method achieves an average performance improvement of 3.47% on the TCIA dataset and 1.75% on the Cardiac dataset. redCompared with MSU-Net, the DSC of our method improves by about 3.4% on the Cardiac dataset. Meanwhile, there is a 0.89% performance improvement on the TCIA dataset compared to CE-Net.

In summary, the results in Table I can prove that orangethe proposed CoAG-MuSFu has achieved better performances than the state-of-the-art baselins in medical image segmentation tasks, and its superior performance is because of the following reasons: 1) compared with single multi-scale information extraction, we combine multi-scale information with extensive semantic information extracted by attention module, which can achieve more effective feature modeling; 2) compared with traditional pure convolutional attention information, CoAG-MuSFu gathers complementary information from adjacent features to generate feature reorganization for content perception, and applies channel weighting to enhance semantic consistency.

Furthermore, the visualized results in Fig. 3 show that our method is superior to the SOTA baselines in terms of accuracy and target morphology capture performance, which benefits from reasonable attention combination and multi-scale morphology capture module. Due to the compression of adjacent organs or other reasons, the CT slices in the 3^{rd} and 4^{th} columns of Fig. 3 show the target discontinuity. When other models incorrectly segment the target into a whole, our method can recognize this situation more accurately and segment it with high precision. In summary, these results further demonstrate that the channel dependencies and long-range spatial information are essential to model the feature representation orangein medical image segmentation.

red

F. Ablation Study

We further investigate the effectiveness and necessity of the proposed modules of EC-CaM-UNet by ablation studies. Specifically, there are five improved components in the proposed EC-CaM-UNet: (i) Channel-Correlation attention module (Abbreviated as CC) and Spatial-Correlation attention module (Abbreviated as SC) are used to respectively obtain spatial and channel correlation attention information, which integrate together will result in the multi-dimensional Collaborative Attention (CoA) Block; (ii) Atrous Spatial Pyramid Pooling Multi-Scale (abbreviated as MuS) Block is used to obtain multi-scale features, which can be combined with CoA block to obtain the proposed CoAG-MuSFu module; (iii) We also introduce Enhanced Convolution (abbreviated as EC) block in the encoder and improve the decoder with side output based deep supervision (abbreviated as DWS) to enhance the feature learning capability of EC-CaM-UNet. The results of ablation studies are shown in Table II.

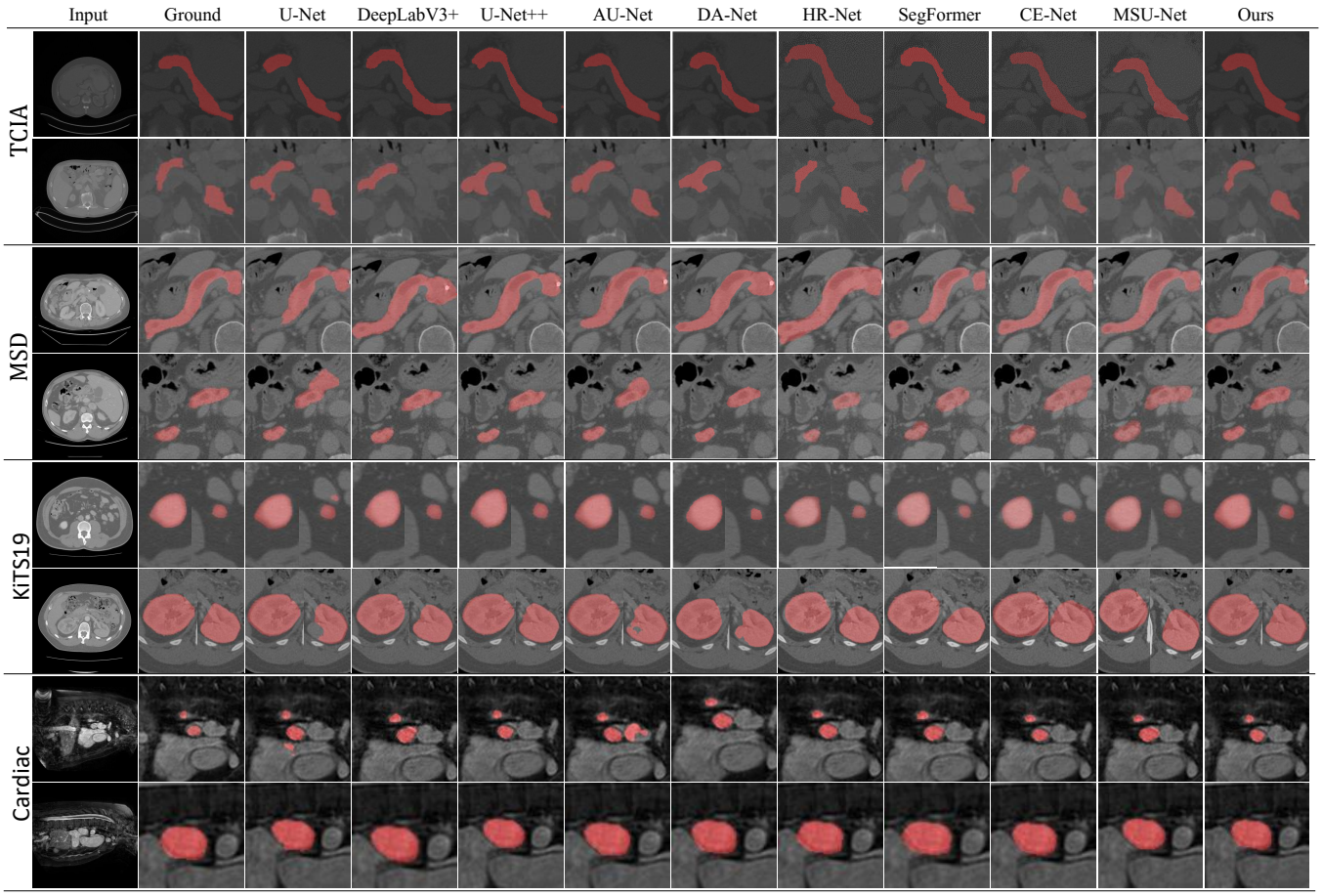


Fig. 3: orangeVisualized segmentation results of our work and the state-of-the-art baselines.

TABLE II: redAblation studies on four datasets, where CC is the Channel-Correlation attention module, SC is Spatial-Correlation attention module, MuS is atrous spatial pyramid polling Multi-Scale (MuS) module, EC is the Enhanced Convolution block, and DWS is the use of decoder with side output to achieve deep supervision.

Method	Module					DATASET											
						TCIA			MSD			KiTS			orangeCardiac		
	CC	SC	MuS	EC	DWS	DSC	Sens	PPV	DSC	Sens	PPV	DSC	Sens	PPV	DSC	Sens	PPV
U-Net						67.80	64.92	81.44	65.35	70.74	77.56	93.83	96.20	94.60	74.41	75.86	78.03
	✓					68.05	65.92	81.07	69.67	72.26	78.03	94.49	96.13	94.86	75.79	77.95	79.00
		✓				68.71	67.14	82.12	67.92	72.36	76.73	94.19	95.99	94.38	77.31	80.04	77.97
			✓			68.14	67.23	80.50	69.68	74.50	75.94	93.42	95.67	93.74	76.98	80.17	77.76
				✓		70.70	67.23	80.51	70.22	74.11	78.71	94.35	96.10	94.59	77.63	79.72	80.19
					✓	69.15	67.01	80.03	69.81	73.05	77.84	94.27	96.13	94.72	76.83	78.55	80.32
	red✓	red✓				red 69.58	red69.24	red81.65	red70.21	red74.50	red76.50	red94.71	red95.34	red94.25	red77.52	red78.36	red80.83
	red✓	red✓	red✓			red71.45	red69.23	red80.14	red70.28	red73.32	red77.29	red93.80	red94.06	red93.10	red78.67	red79.32	red80.53
	✓	✓	✓	✓		71.65	71.98	80.58	71.06	73.65	77.36	94.90	96.19	95.14	78.89	79.85	82.24
Ours	✓	✓	✓	✓	✓	**72.64**	**72.07**	**82.34**	**71.68**	**74.96**	**79.06**	**95.02**	**96.28**	**95.29**	**79.32**	**80.31**	**82.51**

red

1) Effectiveness of Using the Proposed Modules Independently: We first add the five proposed modules, CC, SC, MuS, EC, and DWS, one by one to U-Net, and the results are shown in the 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, and 6^{th} lines of Table II, respectively. By comparing the results of these five intermediate models with those of U-Net, we can find that these five proposed modules all improve the performances of U-Net in terms of all metrics on all four datasets. This thus demonstrates the effectiveness of using these modules in medical image segmentation tasks.

red

2) Effectiveness of Collaborative Attention Guided Multi-Scale Feature Fusion (CoAG-MuSFu): Furthermore, we integrate CC and SC blocks to obtain an intermediate model with multi-dimensional Collaborative Attention (CoA) Block, and further integrate CoA with MuS to obtain another intermediate model with Collaborative Attention Guided Multi-Scale Feature Fusion (CoAG-MuSFu) module; the results of these two intermediate models are shown in the 7^{th} and 8^{th} lines of Table II. By comparing the results of U-Net+CoAG-MuSFu (8^{th} line) with those of U-Net+CoA (7^{th} line) and U-Net+MuS

TABLE III: cyanSegmentation performances of different ways of using the attention module and the multi-scale module.

Dataset	Different Usage Ways	Dice%
TCIA	$CoA \rightarrow MuS$ [8], [10]	69.58
	$CoA + MuS$ [14]	70.13
	Ours	**72.64**
MSD	$CoA \rightarrow MuS$ [8], [10]	70.91
	$CoA + MuS$ [14]	71.22
	Ours	**71.68**
KiTS	$CoA \rightarrow MuS$ [8], [10]	94.64
	$CoA + MuS$ [14]	94.78
	Ours	**95.02**
Cardiac	$CoA \rightarrow MuS$ [8], [10]	77.19
	$CoA + MuS$ [14]	78.02
	Ours	**79.32**

TABLE IV: blueTraining U-Net with different loss functions.

Dataset	Loss	Dice%
TCIA	$\mathcal{L}_{Dice}$	67.80
	$\mathcal{L}_{BCE}$	65.77
	blue$\mathcal{L}_{Dice}$+$\mathcal{L}_{BCE}$	blue68.02
MSD	$\mathcal{L}_{Dice}$	65.35
	$\mathcal{L}_{BCE}$	65.85
	blue$\mathcal{L}_{Dice}$+$\mathcal{L}_{BCE}$	blue66.47
KiTS	$\mathcal{L}_{Dice}$	93.83
	$\mathcal{L}_{BCE}$	93.66
	blue$\mathcal{L}_{Dice}$+$\mathcal{L}_{BCE}$	blue93.89
orangeCardiac	$\mathcal{L}_{Dice}$	74.41
	$\mathcal{L}_{BCE}$	73.13
	blue$\mathcal{L}_{Dice}$+$\mathcal{L}_{BCE}$	blue75.03

(4^{th} line), we can find that U-Net+CoAG-MuSFu constantly outperforms U-Net+CoA and U-Net+MuS in all cases. This thus proves that by using CoA to assign different importance weights to multi-scale features generated in MuS, the proposed CoAG-MuSFu can better highlight important multi-scale features and suppress irrelevant ones to achieve better feature learning, and thus demonstrates our argument that fusing the attention mechanism with multi-scale mechanism (as in the proposed CoAG-MuSFu module) will achieve much better medical image segmentation performances than using them independently (as in the existing attention and/or multi-scale based works).

red

3) Effectiveness of Enhanced Convolution and Decoder with Side Output: Finally, we incorporate the Enhanced Convolution (EC) block into the intermediate model U-Net+CoAG-MuSFu, and then further add side outputs into the model's decoder (i.e., resulting in the proposed EC-CaM-UNet); the results are shown in the last two lines of Table II. By comparing these results with those of U-Net+CoAG-MuSFu (8^{th} line), we can find that the segmentation performances gradually increase with the addition of EC and DWS blocks. Therefore, this proves that integrating EC and DWS blocks with CoAG-MuSFu is also beneficial and essential for the proposed EC-CaM-UNet to achieve the superior medical image segmentation performances.

G. Additional Experiments

1) cyanDifferent Ways of Using Attention and Multi-Scale Modules: cyanAs introduced in Introduction, the most innovative part of EC-CaM-UNet is to propose a CoAG-MuSFu module that can deeply fuse the attention maps generated by the attention module with the multi-scale feature maps generated by the multi-scale module using a 1x1 convolution block; therefore, we conduct additional experiments to prove that combining the attention module and multi-scale module in the proposed deep fusion way is more effective than using them independently in a series way as in [8], [10] or obtaining them simultaneously and then simply adding them together as in [14].

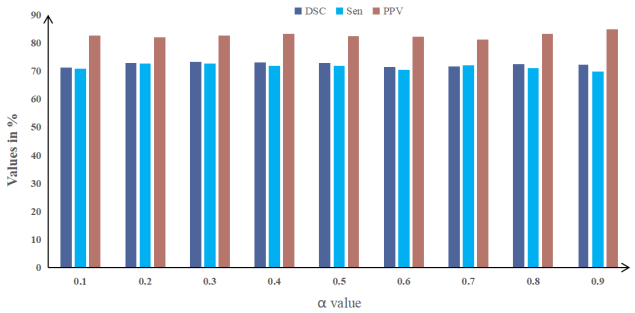


Fig. 4: cyanThe effect of using different weights in the hybrid loss function.

cyanThe results in Table III shows the following observations: (i) the way of obtaining the attention maps and the multi-scale feature maps in parallel and adding them together as in [14] (denoted $CoA \rightarrow MuS$) is better than using attention module and multi-scale module independently in a series way as in [8], [10] (denoted $CoA + MuS$); (ii) our proposed way of deeply fuse them using a 1x1 convolution (denoted Ours) always achieves much better segmentation performances than another two ways. Consequently, these findings well demonstrate the significance and contribution of the proposed CoAG-MuSFu module in finding a new way of using the attention and the multi-scale modules to achieve better performance in medical image segmentation tasks.

2) blueComparison with Different Loss Function: blueTo justify the need of using a hybrid learning loss, we conduct experimental studies to compare the segmentation performances of training U-Net with different loss functions. As shown in Table IV, using solely $\mathcal{L}_{\text{Dice}}$ or $\mathcal{L}_{\text{BCE}}$ can not yield optimal performance; this may be because solely relying on $\mathcal{L}_{\text{Dice}}$ may result in gradient vanishing, and using only $\mathcal{L}_{\text{BCE}}$ may be sensitive to noise samples. Therefore, we use a hybrid loss function, including both $\mathcal{L}_{\text{BCE}}$ and $\mathcal{L}_{\text{Dice}}$ losses, to help model extract more informative and representative features and obtain better medical image segmentation results.

3) cyanCombining Losses in a Weighted Way: cyanWe conduct additional experiments to investigate the effect of combing the losses in a weighted way. Specifically, we redefine the hybrid loss as follows:

$$cyan\mathcal{L}_{\text{Hybrid}} = cyan\alpha \times \mathcal{L}\text{Dice} + (1-\alpha) \times \mathcal{L}_{\text{BCE}}, \quad (17)$$

TABLE V: The segmentation performances of the proposed CoA, its two components (i.e., CC and SC blocks), and the existing SE and CBAM attention modules, where all attention modules are integrated into the end of encoder in U-Net.

Method	DATASET											
	TCIA			MSD			KiTS			orangeCardiac		
	DSC	Sensitivity	Precision	DSC	Sens	Precision	DSC	Sens	Precision	DSC	Sens	Precision
U-Net+SE [25]	68.00	66.82	79.81	66.46	72.61	71.96	94.08	96.71	93.73	75.22	78.36	78.09
U-Net+CBAM [37]	68.16	67.70	78.98	67.49	69.41	72.90	94.19	96.45	93.82	75.52	80.14	75.67
U-Net+CC	68.05	65.92	81.07	69.67	72.26	75.80	94.49	96.13	94.86	76.06	78.10	79.01
U-Net+SC	68.71	67.14	**82.12**	67.92	72.36	**76.73**	94.19	95.99	94.38	76.27	80.94	76.76
U-Net+CoA	**70.24**	**68.68**	82.10	**70.15**	**73.83**	76.24	**94.97**	**96.55**	**95.41**	**79.02**	**80.97**	**80.74**

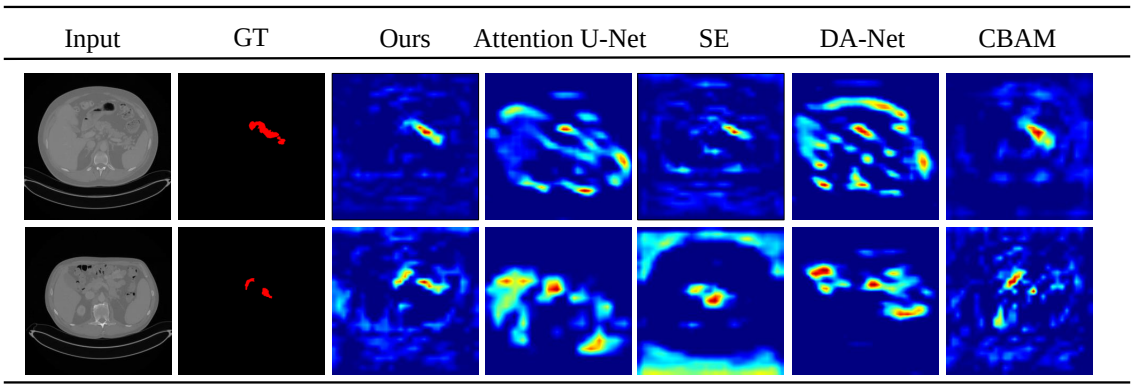


Fig. 5: orangeVisualized attention activate maps on the TCIA dataset, where images from left to right are input images, ground truths, and activate maps of U-Net+CoA (Ours), Attention U-Net, U-Net+SE, DA-Net, and U-Net+CBAM, respectively, and the level of activation gradually increases from blue to red.

cyanwhere α is the weight hyper-parameter.

cyanTo investigate the sensitivity of α on the results, we conducted various experiments on the TCIA dataset, varying α from 0.1 to 0.9. As shown in Fig. 4, the performance of the proposed method is relatively stable (with very limited fluctuations) with the changes of α. For instance, the best results are achieved when α reaches 0.3, with a Dice value of 73.14%, which are only slightly higher than the results obtained when $\alpha = 0.5$ (i.e., the case of adding the losses directly without weights)). Therefore, in this work, the hybrid loss function is constructed by directly adding the losses with the same weight to avoid the very expensive additional time-cost of tuning the hyper-parameter α. However, one can still use the weighted form if very accurate segmentation performances are needed.

4) blueComparison of Attention Mechanisms: We conduct additional experiments to further investigate the effectiveness of the proposed Collaborative Attention (CoA), the two components of CoA, i.e., spatial-correlation (SC) and channel-correlation (CC) attention, and the state-of-the-art attention modules, SE [25] and CBAM [37], where all attention modules are integrated into the end of encoder in U-Net (i.e., the same as in our work). As shown in Table V, using CC attention or SC attention can achieve generally better medical image segmentation performance than using the exciting SE and CBAM attentions. Furthermore, by integrating CC with SC, the proposed CoA achieves the best results, which thus proves the effectiveness of the proposed CoA in our work.

orangeIn Fig. 5, we further visually analyze the heatmaps of U-Net+CoA (ours), Attention U-Net, U-Net+SE, DA-Net, and U-Net+CBAM (as the attention of SegFormer is achieved using Transformer instead of U-Net, we omit it here to keep fair comparison). By comparing the resulting activated maps of these attention methods with ground-truth (GT), we can find that our proposed CoA can capture salient features and highlight relevant target regions with higher intensity in the heatmap than the baselines do. This is consistent with our goal of accurately identifying and segmenting objects of interest in medical images. The success of our proposed attention module is mainly because of the synergistic fusion of attention mechanisms and multi-scale mechanisms; this fusion empowers our model to effectively exploit both local and global information, adaptively allocate attention, and accurately delineate the regions of interest. The clear distinction and localization of regions of interest in the heatmap reaffirm the effectiveness of our method in accurately highlighting important structures and aiding in the precise segmentation process.

V. BLUEDISCUSSION AND FUTURE WORK

This section will introduce the main differences between our work and the existing research methods, the social impact of the proposed method, and the limitations and future works.

Differences between our work and the existing research methods. redThe main contribution is to propose a collaborative attention multi-scale fusion module, CoAG-MuSFu, which is different from the existing multi-scale mechanism [23], [24] by using an attention mechanism to weight the multi-scale

features to highlight key multi-scale features and suppress irrelevant ones. Our method is based on both global and local multi-scale. In this way, the local-to-global morphological feature was efficiently integrated for comprehensive characterization of the pancreas. Additionally, we embedded an attention mechanism module to obtain enhanced connections between feature channels and spatial features. Our proposed attention module is also different from existing module [7], [15], [25]. In the substructure of channel attention, we learned from SE-Net's structure and designed a one-dimension encoder-decoder by 1D convolution to compress and distribute data from different channels while reducing the amount of computation. We extracted the two kinds of attention information separately and applied them in a cascade way. By the fusion of two kinds of attention mechanisms, we not only reduce the redundant convolution caused by parallel feature concatenation but also ensure that the extracted attention information is the original feature without weighting [37].

blue**Social impact.** Our work has the potential to make a significant impact on clinical practices by improving the accuracy and efficiency of medical image segmentation, which is a critical step in the diagnosis and treatment of various medical conditions. For instance, the segmentation accuracy for kidney organs has reached a high level, enabling the direct utilization of the proposed model in clinical applications. The segmented results can be directly provided to physicians for diagnosis and treatment, significantly reducing the cost associated with organ and lesion region assessment and analysis in clinical settings.

blue**Limitations and future works.** Despite demonstrating better performances in medical image segmentation tasks, we observed in our experiments that the improvements of the proposed segmentation model's segmentation performances on MSD and TCIA datasets are not as significant as those on KiTS19 and Cardiac datasets; this may be because, compared to the pancreas, the shapes of kidneys (in KiTS dataset) and hearts (in Cardiac dataset) are relatively regular, and the contrast of its images is relatively high, so the advantages of using our methods may be limited; therefore, it is an interesting future work to discover the solution to further enhance the performances on this kind of datasets. In addition, the proposed method is currently designed to process 2D medical images, so it is also interesting to further extend the proposed model to make it capable of processing 3D medical images in the future works; a potential direction of achieving this may further enhance the dimension of the attention mechanism to make it able to model the dependencies between 3D volumes.

VI. Conclusion

In this study, we proposed a CoAG-MuSFu Module to construct the weighted feature maps and make our network pay more attention to the pancreas that is small and with various shapes. We select two public pancreas datasets and a kidney dataset to evaluate the model's effectiveness. On the challenging pancreas datasets, our method outperformed the current SOTA models, and realized the comparable performance on the easier task of kidney segmentation. This also confirms that our model has a targeted improvement in the segmentation of the pancreas or organs with the related characteristics.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

[2] Qi Bi, Kun Qin, Han Zhang, Zhili Li, and Kai Xu. Radc-net: A residual attention based convolution network for aerial scene classification. *Neurocomputing*, 377:345–359, 2020.

[3] Qi Bi, Han Zhang, and Kun Qin. Multi-scale stacking attention pooling for remote sensing scene classification. *Neurocomputing*, 436:147–161, 2021.

[4] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018.

[5] Xiaohan Ding, Yuchen Guo, Guiguang Ding, and Jungong Han. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1911–1920, 2019.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019.

[8] Chengling Gao, Hailiang Ye, Feilong Cao, Chenglin Wen, Qinghua Zhang, and Feng Zhang. Multiscale fused network with additive channel–spatial attention for image segmentation. *Knowledge Based Systems*, 214:106754, 2021.

[9] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292, 2019.

[10] Pengcheng Guo, Xiangdong Su, Haoran Zhang, and Feilong Bao. Mcdalnet: Multi-scale contextual dual attention learning network for medical image segmentation. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2021.

[11] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 603–612, 2019.

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. pages 448–456, 2015.

[13] Lei Ke, Xia Li, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Prototypical cross-attention networks for multiple object tracking and segmentation. *Advances in Neural Information Processing Systems*, 34:1192–1203, 2021.

[14] Haixing Li, Haibo Luo, Wang Huan, Zelin Shi, Chongnan Yan, Lanbo Wang, Yueming Mu, and Yunpeng Liu. Automatic lumbar spinal mri image segmentation with a multi-scale attention network. *Neural Computing and Applications*, 33:11589–11602, 2021.

[15] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[16] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1):3, 2010.

[17] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[18] Run Su, Deyun Zhang, Jinhuai Liu, and Chuandong Cheng. Msu-net: Multi-scale u-net for 2d medical image segmentation. *Frontiers in Genetics*, 12:639930, 2021.

[19] blueBen-Cohen, Avi and Diamant, Idit and Klang, Eyal and Amitai, Michal and Greenspan, Hayit. blueFully convolutional network for liver segmentation and lesions detection. In *blueDeep Learning and Data Labeling for Medical Applications*, pages blue77–85. blue2016.

[20] blueBrosch, Tom and Tang, Lisa YW and Yoo, Youngjin and Li, David KB and Traboulsee, Anthony and Tam, Roger. blueDeep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *blueIEEE Transactions on Medical Imaging*, blue35(blue5):blue1229–1239, blue2016.

[21] blueChen, Liang-Chieh and Papandreou, George and Kokkinos, Iasonas and Murphy, Kevin and Yuille, Alan L. blueDeeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *blueIEEE Transactions on Pattern Analysis and Machine Intelligence*, blue40(blue4):blue834–848, blue2017.

[22] blueChen, Liang-Chieh and Zhu, Yukun and Papandreou, George and Schroff, Florian and Adam, Hartwig. blueEncoder-decoder with atrous separable convolution for semantic image segmentation. In *blueProceedings of the European Conference on Computer Vision*, pages blue801–818, blue2018.

[23] blueGao, Yunhe and Liu, Chang and Zhao, Liang. blueMulti-resolution Path CNN with Deep Supervision for Intervertebral Disc Localization and Segmentation. In *blueProceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages blue309–317. blueSpringer, blue2019.

[24] blueHe, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. blueSpatial pyramid pooling in deep convolutional networks for visual recognition. *blueIEEE Transactions on Pattern Analysis and Machine Intelligence*, blue37(blue9):blue1904–1916, blue2015.

[25] blueHu, Jie and Shen, Li and Sun, Gang. blueSqueeze-and-excitation networks. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue7132–7141, blue2018.

[26] blueHuang, Huimin and Lin, Lanfen and Tong, Ruofeng and Hu, Hongjie and Zhang, Qiaowei and Iwamoto, Yutaro and Han, Xianhua and Chen, Yen-Wei and Wu, Jian. blueUnet 3+: A full-scale connected unet for medical image segmentation. In *blueProceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages blue1055–1059. blueIEEE, blue2020.

[27] blueJha, Debesh and Smedsrud, Pia H and Riegler, Michael A and Johansen, Dag and De Lange, Thomas and Halvorsen, Pål and Johansen, Håvard D. blueResunet++: An advanced architecture for medical image segmentation. In *blueIEEE International Symposium on Multimedia*, pages blue225–2255. blueIEEE, blue2019.

[28] blueLin, Tsung-Yi and Dollár, Piotr and Girshick, Ross and He, Kaiming and Hariharan, Bharath and Belongie, Serge. blueFeature pyramid networks for object detection. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue2117–2125, blue2017.

[29] blueLong, Jonathan and Shelhamer, Evan and Darrell, Trevor. blueFully convolutional networks for semantic segmentation. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue3431–3440, blue2015.

[30] bluePham, Dzung L and Xu, Chenyang and Prince, Jerry L. blueCurrent methods in medical image segmentation. *blueAnnual Review of Biomedical Engineering*, blue2(blue1):blue315–337, blue2000.

[31] blueQi, Kehan and Yang, Hao and Li, Cheng and Liu, Zaiyi and Wang, Meiyun and Liu, Qiegen and Wang, Shanshan. blueX-net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies. In *blueProceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages blue247–255. blueSpringer, blue2019.

[32] blueRickmann, Anne-Marie and Roy, Abhijit Guha and Sarasua, Ignacio and Navab, Nassir and Wachinger, Christian. blue'Project & Excite'Modules for Segmentation of Volumetric Medical Scans. In *blueProceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*, pages blue39–47. blueSpringer, blue2019.

[33] blueRonneberger, Olaf and Fischer, Philipp and Brox, Thomas. blueU-net: Convolutional networks for biomedical image segmentation. In *blueProceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages blue234–241. blueSpringer, blue2015.

[34] blueWang, Bo and Qiu, Shuang and He, Huiguang. blueDual encoding u-net for retinal vessel segmentation. In *blueProceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages blue84–92. blueSpringer, blue2019.

[35] blueWang, Lei and Wang, Bo and Xu, Zhenghua. blueTumor Segmentation Based on Deeply Supervised Multi-Scale U-Net. In *blueProceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pages blue746–749. blueIEEE, blue2019.

[36] blueWang, Xiaolong and Girshick, Ross and Gupta, Abhinav and He, Kaiming. blueNon-local neural networks. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue7794–7803, blue2018.

[37] blueWoo, Sanghyun and Park, Jongchan and Lee, Joon-Young and So Kweon, In. blueCbam: Convolutional block attention module. In *blueProceedings of the European Conference on Computer Vision*, pages blue3–19, blue2018.

[38] blueYu, Qihang and Xie, Lingxi and Wang, Yan and Zhou, Yuyin and Fishman, Elliot K and Yuille, Alan L. blueRecurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue8280–8289, blue2018.

[39] blueZhao, Hengshuang and Shi, Jianping and Qi, Xiaojuan and Wang, Xiaogang and Jia, Jiaya. bluePyramid scene parsing network. In *blueProceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages blue2881–2890, blue2017.

[40] blueZhou, Xiangrong and Takayama, Ryosuke and Wang, Song and Hara, Takeshi and Fujita, Hiroshi. blueDeep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method. *blueMedical Physics*, blue44(blue10):blue5221–5233, blue2017.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[43] Chang Wang, Zongya Zhao, Qiongqiong Ren, Yongtao Xu, and Yi Yu. Dense u-net based on patch-based learning for retinal vessel segmentation. *Entropy*, 21(2):168, 2019.

[44] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3939–3948, 2019.

[45] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3349–3364, 2020.

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[47] Zhenghua Xu, Tianrun Li, Yunxin Liu, Yuefu Zhan, Junyang Chen, and Thomas Lukasiewicz. PAC-Net: Multi-pathway FPN with position attention guided connections and vertex distance IoU for 3D medical image detection. *Frontiers in Bioengineering and Biotechnology*, 11:1049555, 2023.

[48] Zhenghua Xu, Shijie Liu, Di Yuan, Lei Wang, Junyang Chen, Thomas Lukasiewicz, Zhigang Fu, and Rui Zhang. ω-net: Dual supervised medical image segmentation with multi-dimensional self-attention and diversely-connected multi-scale convolution. *Neurocomputing*, 500:177–190, 2022.

[49] Zhenghua Xu, Chang Qi, and Guizhi Xu. Semi-supervised attention-guided CycleGAN for data augmentation on medical images. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pages 563–568, 2019.

[50] Zhenghua Xu, Xudong Zhang, Hexiang Zhang, Yunxin Liu, Yuefu Zhan, and Thomas Lukasiewicz. Efpn: Effective medical image detection using feature pyramid fusion enhancement. *Computers in Biology and Medicine*, page 107149, 2023.

[51] Dan Yao, Zhenghua Xu, Yi Lin, and Yuefu Zhan. Accurate and intelligent diagnosis of pediatric pneumonia using x-ray images and blood testing data. *Frontiers in Bioengineering and Biotechnology*, 11:1058888, 2023.

[52] Shuo Zhang, Jiaojiao Zhang, Biao Tian, Thomas Lukasiewicz, and Zhenghua Xu. Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation. *Medical Image Analysis*, 83:102656, 2023.

Zhenghua Xu received a M.Phil. in Computer Science from The University of Melbourne, Australia, in 2012, and a D.Phil in computer Science from University of Oxford, United Kingdom, in 2018. From 2017 to 2018, he worked as a research associate at the Department of Computer Science, University of Oxford. He is now a professor at the Hebei University of Technology, China, and a awardee of "100 Talents Plan" of Hebei Province. He has published more than 30 papers in top AI or database conferences and journals, e.g., NeurIPS, AAAI, IJCAI, ICDE, IEEE TNNLS, Medical Image Analysis, IEEE TCS, etc. His current research focuses on intelligent medical image analysis and computer vision.

Biao Tian is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. He received B.Eng. degree in Electrical Engineering and Automatics from City College of Hebei University of Technology, China, in 2020. His research interests lie in medical image processing using deep learning methods.

Shijie Liu is currently a master student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, China. He received B.Eng. degree in Biomedical Engineering from Hebei University of Technology, China, in 2019. His research interests lie in medical image processing using machine learning and deep learning methods.

Xiangtao Wang is currently a master student at Hebei University of Technology, China. He received the B.Eng. degree in Electrical Engineering and Automation from the North China University of Science and Technology, China, in 2021. His research interests lie in image analysis using deep learning methods.

Di Yuan is currently a PhD student in the State Key Laboratory of Reliability and Intelligence of Electrical Equipment, Hebei University of Technology, China. She received B.Eng. degree in Electrical Engineering and Automatics from Tianjin University of Technology and Education, China, in 2016. Her research interests lie in medical image processing using deep learning methods and reinforcement learning.

Junhua Gu received the B.S degree in mathematics from Shanghai Jiaotong University, Shanghai, China, in 1988, the M.S. degree in computer science and the Ph.D. degree in electrical engineering from the Hebei University of Technology, Tianjin, China, in 1993 and 1997, respectively. He is currently a Professor at the Hebei University of Technology, China. He has authored more than 70 papers. His current research interests include big data, intelligent control, and intelligent transportation systems. Prof. Gu was awarded the Hebei New Century "333 Talent Project" Second Level Suitable Person.

Junyang Chen received the Ph.D. degree in computer and information science from the University of Macau, Macau, China, in 2020. He is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. His research interests include graph neural networks, text mining, deep learning, and recommender systems.

Thomas Lukasiewicz is a Professor of Computer Science at the Department of Computer Science, University of Oxford, UK, heading the Intelligent Systems Lab within the Artificial Intelligence and Machine Learning Theme. He currently holds an AXA Chair grant on "Explainable Artificial Intelligence in Healthcare" and a Turing Fellowship at the Alan Turing Institute, London, UK, which is the UK's National Institute for Data Science and Artificial Intelligence. He received the IJCAI-01 Distinguished Paper Award, the AIJ Prominent Paper Award 2013, the RuleML 2015 Best Paper Award, and the ACM PODS Alberto O. Mendelzon Test-of-Time Award 2019. He is a Fellow of the European Association for Artificial Intelligence (EurAI) since 2020. His research interests are especially in artificial intelligence and machine learning.

Victor C. M. Leung (Life Fellow, IEEE) is currently a Distinguished Professor of computer science and software engineering with Shenzhen University, Shenzhen, China. He is also an Emeritus Professor of electrical and computer engineering and the Director of the Laboratory for Wireless Networks and Mobile Systems, University of British Columbia (UBC), Vancouver, BC, Canada. He is a Fellow of Royal Society of Canada, Canadian Academy of Engineering, and Engineering Institute of Canada. He is named in the current Clarivate Analytics list of "Highly Cited Researcher". His research interests include wireless networks and mobile systems.