FCP-Net: A Feature-Compression-Pyramid Network Guided by Game-Theoretic Interactions for Medical Image Segmentation

Yexin Liu, Jian Zhou, Lizhu Liu, Zhengjia Zhan, Yueqiang Hu, YongQing Fu, Huigao Duan

Abstract—**Medical image segmentation is a crucial step in diagnosis and analysis of diseases for clinical applications. Deep neural network methods such as DeepLabv3+ have successfully been applied for medical image segmentation, but multi-level features are seldom integrated seamlessly into different attention mechanisms, and few studies have explored the interactions between medical image segmentation and classification tasks. Herein, we propose a feature-compression-pyramid network (FCP-Net) guided by game-theoretic interactions with a hybrid loss function (HLF) for the medical image segmentation. The proposed approach consists of segmentation branch, classification branch and interaction branch. In the encoding stage, a new strategy is developed for the segmentation branch by applying three modules, e.g., embedded feature ensemble, dilated spatial mapping and channel attention (DSMCA), and branch layer fusion. These modules allow effective extraction of spatial information, efficient identification of spatial correlation among various features, and fully integration of multi-receptive field features from different branches. In the decoding stage, a DSMCA module and a multi-scale feature fusion module are used to establish multiple skip connections for enhancing fusion features. Classification and interaction branches are introduced to explore the potential benefits of the classification information task to the segmentation task. We further explore the interactions of segmentation and classification branches from a game theoretic view, and design an HLF. Based on this HLF, the segmentation, classification and interaction branches can collaboratively learn and teach each other throughout the training process, thus applying the conjoint information between the segmentation and classification tasks and improving the generalization performance. The proposed model has been evaluated using several datasets, including ISIC2017, ISIC2018, REFUGE, Kvasir-SEG, BUSI, and PH2, and the results prove its competitiveness compared with other state-of-the-art techniques.**

***Index Terms*—Hybrid loss function, game theory, embedded feature ensemble module, dilated spatial mapping and channel attention module, branch layer fusion module.**

I. INTRODUCTION

MEDICAL image segmentation is an important step in medical image analysis, such as skin lesion segmentation, optic disc segmentation, polyp segmentation, and breast ultrasound image segmentation. The conventional approaches for medical image segmentation usually employ low-level attributes that are solely relied on pixel-level features. As such, they often cannot achieve satisfactory diagnostic performance because of the low contrast or surface artifacts [5].

Methods based on deep convolutional neural network (CNN) have previously been applied in the medical image segmentation. Most of them are based on fully connected networks (FCNs) and U-Net. These powerful algorithms have the abilities to capture prominent context information from images, rather than hand-crafted features, making the network to be robust to local image transformations [7], [10]. However, these methods typically employ small kernels, which unavoidably leads to a loss of localization or a decreased spatial feature resolution [12]. In addition, the objects to be segmented have varied scales, thus prohibiting an accurate segmentation [9]. To solve these problems, many models were proposed to integrate multi-scale context information or leverage global context for an effective segmentation [13]. Xue et al. (2018) proposed an adversarial network with multi-scale l1 loss for medical image segmentation to capture the multi-scale information [14]. Chen et al. (2018) proposed the use of DeepLabv3+ as a simple yet effective decoder module for sharp segmentation [3].

Recently, attention-based networks have been widely applied in medical image segmentation, because the attention strategy avoids the uses of multiple and similar feature maps in the network and focuses only on the most informative features for a given task without any additional supervisions [15], [16], [17] For example, Sarker et al. (2020) proposed a model of lightweight generative adversarial networks that combines 1-D kernel factorized networks, position and channel attention, and multiscale aggregation mechanisms, thus achieving a precise skin lesion segmentation with minimum resources [18]. Azad et al. (2020) proposed attention DeepLabv3+ to focus on capturing more relevant features by employing attention mechanisms in two stages [9].

Manuscript submitted June 9, 2021. (Corresponding author: Jian Zhou, jianzhou@hnu.edu.cn)

Yexin Liu, Jian Zhou, Lizhu Liu, Zhengjia Zhan, Y. Hu and Huigao Duan are with Engineering Research Center of Automotive Electrics and Control Technology, College of Mechanical and Vehicle Engineering, Hunan University, Changsha 410082, China (e-mail: lyx1223@hnu.edu.cn; jianzhou@hnu.edu.cn; liulz@hnu.edu.cn; 1369391193@qq.com; huyq@hnu.edu.cn; duanhg@hnu.edu.cn). Richard Yongqing Fu is with Faculty of Engineering and Environment, Northumbria University at Newcastle, UK, NE1 8ST (email: richard.fu@northumbria.ac.uk).

In addition, multi-task learning is often utilized in medical image segmentation, as the segmentation and classification are highly correlated. Applying the joint information between classification and segmentation tasks can enhance the performance of segmentation tasks. Chen et al. (2018) designed a feature passing module to pass messages between skin lesion segmentation branch and classification branch to make a full use of features from different tasks [19]. Xie et al. (2020) proposed a two-step mutual bootstrapping deep convolutional neural network for skin lesion analysis [35]. He et al. (2020) used multi-task learning to determine the contour of organs-at-risk in CT images [20].

Although these methods have achieved remarkable success, there still exist many limitations. Firstly, the stability of attention weights and the complementariness between attention mechanisms and residual blocks have not been sufficiently exploited to mitigate the challenges of lesion segmentation. Secondly, multi-level features are seldom seamlessly integrated into different attention mechanisms, which may lead to a redundant use of low-level features. Finally, in order to learn additional feature representations and improve the generalization ability of the model, previous multi-task learning methods usually use feature pass modules or two-step training strategies, which increase the parameters of the model and the complexity of inferences.

Herein, inspired by the success of the aforementioned deep models, especially the DeepLabv3+, we propose a feature-compression-pyramid network (FCP-Net) guided by game-theoretic interactions with a hybrid loss function (HLF) to address the above challenges. The proposed FCP-Net consists of three branches, e.g., segmentation (the main branch), classification, and interaction branches. Compared with the DeepLabv3+, in the segmentation branch and at the encoding stage, a new strategy is proposed by designing modules of embedded feature ensembles (EFE), dilated spatial mapping and channel attention (DSMCA), and branch layer fusion (BLF), in order to extract spatial, channel, and multi-scale information of the target. Inspired by the squeeze-and-excitation network (SE-Net) [21], an EFE module is employed to (1) adaptively capture the explicit relationships among the channels in convolution layers, (2) focus on informative features, and (3) suppress the unnecessary characteristics by utilizing global information from the input data. A DSMCA module is developed to merge multi-scale contextual and channel information, by integrating features from different receptive fields. As a result, this system is capable of learning nonlinear interactions among different channels. The BLF module is used to aggregate features extracted in different branches by the EFE. This approach combines information from different receptive fields with global information from the images. At the decoding stage, the DSMCA and multi-scale feature fusion (MSFF) modules are used to integrate information from featured maps of different resolutions, combining features from different stages, in order to improve the fusion between the encoder and decoder.

In addition, classification branch and interaction branch are introduced to explore the potential benefit of the classification information task to the segmentation task. The interaction branch only exists in the training stage, which reduces the complexity of model inference. Furthermore, we explore the interactions of medical image segmentation and classification from a game theoretic view, and design a hybrid loss function according to the game-theoretic interactions. Based on this hybrid loss function, the segmentation, classification and interaction branches can collaboratively learn and teach each other throughout the training process, thus applying the conjoint information between segmentation and classification tasks and improving the generalization performance.

To the best of our knowledge, this is the first work using the interactions between related tasks from a game theoretic view, in order to improve the performance of the convolution neural network (CNN) for medical image segmentation.

Our key contributions are summarized as follows:

(1) In the segmentation branch, at the encoding stage, three novel modules including EFE, DSMCA, and BLF are proposed to effectively capture context information and fuse multi-scale features. At the encoding stage, a strategy combing the DSMCA and MSFF modules is employed to fuse the features of different resolutions.

(2) An HLF is proposed from the game theoretic view. Based on this interaction, different branches can collaboratively learn and teach each other throughout the training process, thus applying the conjoint information between segmentation and classification tasks and improving the generalization performance.

(3) The proposed model demonstrates its successful applications with state-of-the-art performance across several different datasets (ISIC2017, ISIC2018, REFUGE, Kvasir-SEG, BUSI, and PH2) and is proven to be a promising strategy for various medical image segmentation tasks across multiple modalities.

II. RELATED WORK

A. CNNs for medical image segmentation

FCNs were among the first groups of models to be trained for end-to-end and pixel-wise predictions [22]. However, small size objects in the images were often ignored or incorrectly classified as the background by these algorithms. Motivated by the FCN architecture, Ronneberger et al. proposed the U-Net, which consists of encoding and decoding paths and yields highly precise segmentation results with a few training samples [1]. However, the U-Net has its limitations for explicitly modeling long-range dependency, due to its intrinsic locality of convolution operations. Motivated by the U-Net and Inception-ResNet structures [23], [24], Gu et al. proposed a context encoder network (CE-Net) to capture high-level information and preserve spatial information for medical image segmentation tasks, such as optic disc segmentation and blood vessel detection [6]. However, the CE-Net cannot dynamically adjust the receptive fields to fit the targets with different sizes. Recently, many scale-aware methods based on the attention mechanisms have been developed to mitigate the above problems. Qin et al. proposed an autofocus layer to adaptively change the size of effective receptive field based on the processed context to generate more powerful features [25]. However, the stability of attention weights and the complementariness between attention mechanisms and residual

blocks have not been sufficiently exploited. In our wok, we developed an EFE module, aiming at adaptively capturing the explicit relationships among the channels in convolution layers.

B. Multi-scale learning

To mitigate the problems of conventional CNN that ignores small size objects in the images or incorrectly classifies them as the background, many models were proposed to integrate multi-scale context information and leverage more global context for an effective segmentation [13]. Pyramid scene parsing network (PSPNet) [26] exploits global context information for region-based aggregation. DeepLab model combines several paralleled atrous convolutions (e.g., a process called atrous spatial pyramid pooling) to integrate multi-scale information [27]. SLSDeep model combines skip-connections, dilated residual and pyramid pooling networks and formulates a new loss function to accurately segment the boundaries of melanoma regions [28]. Inspired by the generatively adversarial networks , Xue et al. (2018) proposed an end-to-end adversarial neural network (SegAN) and a multi-scale loss function to learn semantic features of skin lesions, finding it more effective for the segmentation task and more stable in training [29]. However, in each single stage of these methods, there are seldom effective extraction and utilization of multi-scale context information, which may cause a redundant use of low-level features. To effectively extract multi-scale features, , we employ a multi-branch architecture and propose a DSMCA module, which integrates channel attention and two context attentions with different receptive fields to extend the receptive field of the network and capture multi-scale features.

C. Multi task learning

Multi-task learning (MTL) is a general method for improving generalization by learning tasks in parallel. In recent years, the MTL has been deployed for medical image segmentation and classification. Some methods utilized the lesion segmentation results to filter the distractions, and thus improved the classification performance [30], [31], [32]. Whereas the others explored the potential benefit of classification results to the lesion segmentation task. For example, Hong et al. [33] proposed a decoupled network for weakly-supervised segmentation, where the class-specific activation maps are transferred from the classification network to the segmentation network. Xie et al. [34] proposed a mutual bootstrapping deep convolutional neural network model for simultaneous skin lesion segmentation and classification. Vandenhende et al. proposed a multiscale multi-modal distillation unit, a feature propagation module, and a feature aggregation unit to model the task at different scales [35]. However, this may cause the increase of parameters and require much longer running time for model inferences, which is often difficult to be deployed in the clinical settings.

Recently, interactions have been extensively investigated in the task of interpreting neural networks. Game theory is one of the representative techniques to be used. For example, Zhang et al. (2020) improved the utility of dropout by proposing a loss function based on the game theory [36]. However, game-theoretic interactions have never been explored for multi-task learning in the medical image segmentation field.

In this work, motivated by game-theoretic interactions [36], we explore the interactions between medical image segmentation and classification from a game theoretic view and design an HLF based on the game-theoretic interactions. As mentioned above, compared with the previous methods which generally design a feature passing module to pass messages, which increase the parameters of the model and the complexity of inference, our new method can make the branches of segmentation, classification and interaction collaboratively learn and teach each other throughout the training process. In addition, because the designed interaction branch only exists in the training stage, the parameters of the model can be decreased and the complexity of model inferences can be reduced.

III. Proposed Method

A. Overview

The proposed FCP-Net consists of three branches, e.g., segmentation, classification, and interaction (see Fig. 1). The main branch is the segmentation branch for medical image segmentation. The classification branch and interaction branch are the auxiliary branches, which are utilized to explore the interactions between the segmentation and classification. The implementation details are provided below.

In the segmentation branch, an encoder network is utilized for learning latent representations of input data and a decoder network is used for reconstructing information from the encoder module. Here, a new strategy is proposed for the encoding stage and includes the modules of EFE, DSMCA, and BLF. These three components are used to: (1) adaptively recalibrate feature responses based on contextual information and weights; (2) capture the spatial correlations among different features and focus attention on the channel relationships to improve performance; and (3) integrate multi-branch information for the target, respectively. The EFE module is employed to adaptively capture the explicit relationships among the convolutional layer channels. This is performed to focus on useful features and also suppress the others, by utilizing a light weight attention mechanism. The DSMCA module is proposed to effectively extract and utilize the multi-scale context information. Finally, the BLF module is used to integrate multi-receptive field information among the different branches. At the decoding stage, the DSMCA and multi-scale feature fusion (MSFF) modules are used to establish multiple skip connections and improve the contextual information fusion between the encoder and decoder.

A classification branch is added to the end of the encoder network for the segmentation branch. The classification branch contains a global average pooling layer, two fully connected layer, and an activation function, which can predict the input image to be benign or malignant. The interaction branch provides interaction information between classification and segmentation branches. As explained before, the interaction branch is only used in the training stage.

B. An embedded feature ensemble for separable convolution module

Five branches of ASPP in a DeepLabv3+ network are directly concatenated and passed to the decoder without further

extraction or integration of the information [3]. The EFE module, inspired by an SE network [21], is proposed to capture the channel information (see Fig. 3).

Unlike in the SE block, the activation function of a rectified

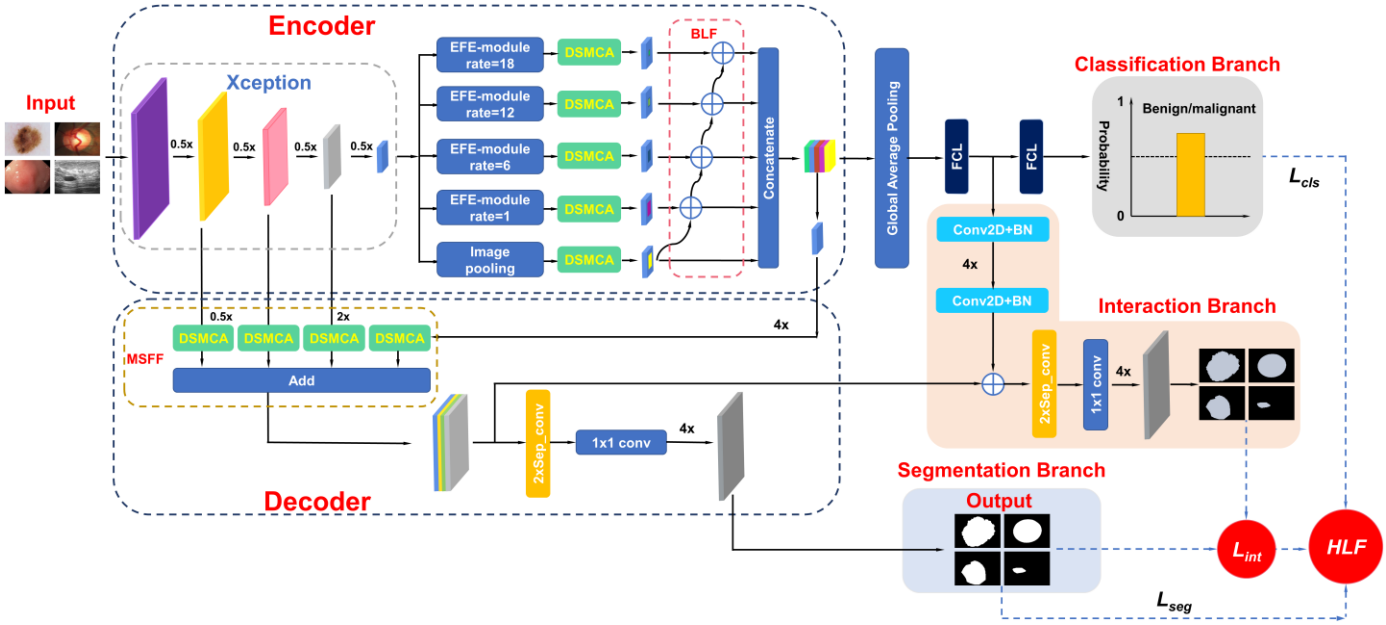

Figure 1. The FCP network architecture. EFE: embedded feature ensemble for separable convolution. DSMCA: dilated spatial mapping and channel attention. BLF: branch layer fusion. MSFF: multi-scale feature fusion. FCL: fully connected layer. L_{cls}: loss function of skin lesion classification branch. L_{seg}: loss function of skin lesion segmentation branch. L_{int}: interaction loss of skin lesion classification and segmentation.

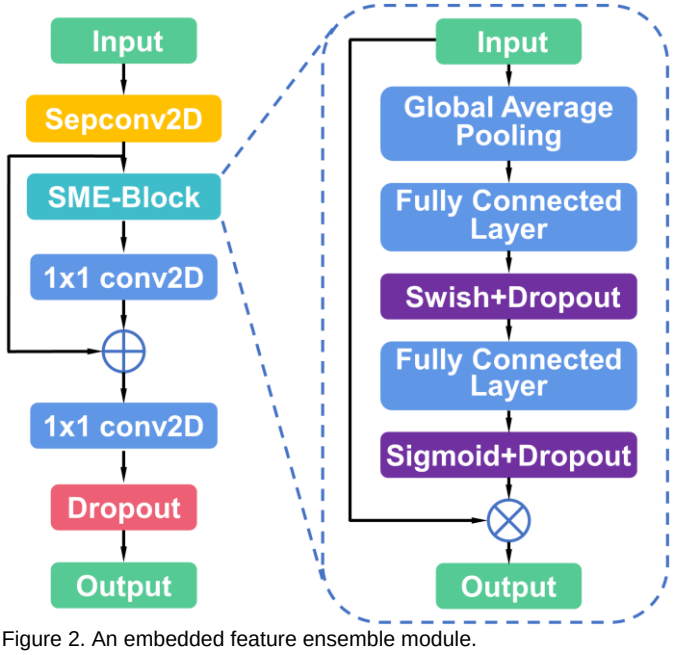

Figure 2. An embedded feature ensemble module.

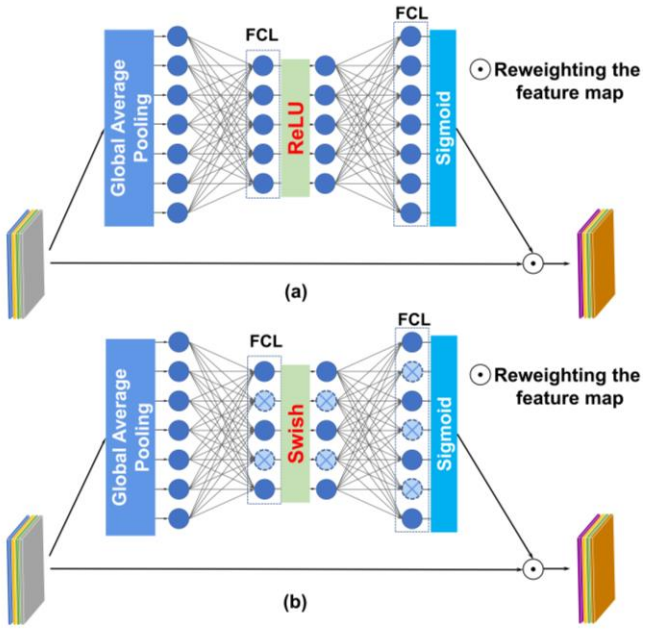

Figure 3. FCL: fully connected layer. (a) Squeeze and excitation in SE-block. (b) Squeeze and modified excitation in SME-block. Dropout was applied in fully connected layers. ReLU is replaced with Swish function.

the relationships among convolution layer channels via an attention mechanism and a residual block (see Fig. 2).

Depth-wise separable convolutions (with a given dilation rate) are employed to project depth-wise channels in the first component of the EFE module [37]. The outputs of the depth-wise separable convolution layer are then transmitted to the squeeze and multi-excitation (SME) blocks. The squeeze operation is used to extract contextual information outside the local region by conducting a global average pooling. A modified excitation operation is proposed to adaptively capture

linear unit (ReLU) is replaced with a Swish (which is a nonlinear term used as a drop-in replacement for the activation function) in the excitation operation. The nonlinearity of this operation is defined as $x \times \mathrm{sigmoid}(\beta x)$, where x is the input of the activation function and β is a constant or trainable parameter. This operation has been applied to demonstrate the significant improvement of the accuracy of neural networks

[38]. Dropout layers are then added after the fully connected layer, and used to prevent complex co-adaptations among the different channels, thereby promoting each channel to encode the useful information itself. The mathematical formulation of the SME block can be represented as:

$$z = \frac{1}{H \cdot W} \cdot \sum_{i=1}^{H} \sum_{j=1}^{W} Sep(X_{input}, rate) \quad (1)$$

$$s = \sigma(Dropout(W_2 \cdot Dropout(\delta(W_1 \cdot z)))) \quad (2)$$

$$x_i = s_i \cdot Sep_i \quad (3)$$

$$X = [x_1, x_2, \cdots, x_i, \cdots, x_c] \quad (4)$$

where X_{input} denotes the input of SME block. H and W are the height and width of inputs, respectively. Sep is the output of the depth-wise separable convolution, rate is the dilation rate of depth-wise separable convolution, z is the output of the global average pooling step, W_1 and W_2 are the parameters of the two fully connected layers, δ and σ are the Swish and Sigmoid functions, respectively, s is the learned scale factor and s_i is the learned scale factor for the i^{th} channel, x_i is the output of the i^{th} channel of the SME block and X is the output of the SME block.

The non-exclusive relationships are then further obtained between the depth-wise separable convolution and the SME block, using a residual connection and a 1×1 convolution layer. The output of these layers can be represented as:

$$U = Dropout(W_4 \cdot (Sep + W_3 \cdot X)) \quad (5)$$

where W_3 and W_4 are the parameters of the two convolution layers and U is the output of the EFE module.

C. Dilated spatial mapping and channel attention modules

A second module is developed to capture the contextual information and extract useful features. This is aimed to mitigate the challenges of multi-level features which have not been seamlessly integrated into different attention mechanisms. To fulfil this objective, the module should meet three criteria. Firstly, a multi-scale fusion approach is required, being capable of integrating multi-scale information and learning nonlinear interactions of the aggregated information. Secondly, the module should adaptively combine local features with their global dependencies [4]. Finally, the module should require as few parameters as possible. A structure combining channel and multi-receptive field spatial attention mechanisms is employed to meet these criteria, as shown in Fig. 4. The mathematical formula to describe this dilated spatial mapping and channel attention (DSMCA) module are as follows:

$$S = SME(X_{in}) \quad (6)$$

$$DSM\_B_1 = X_{in} \cdot \sigma(W_8 \cdot (Dilation\_Conv(W_7 \cdot X_{in}, r = 2))) \quad (7)$$

$$DSM\_B_2 = X_{in} \cdot \sigma(W_{10} \cdot (Dilation\_Conv(W_9 \cdot X_{in}, r = 4))) \quad (8)$$

$$O = S + DSM\_B_1 + DSM\_B_2 \quad (9)$$

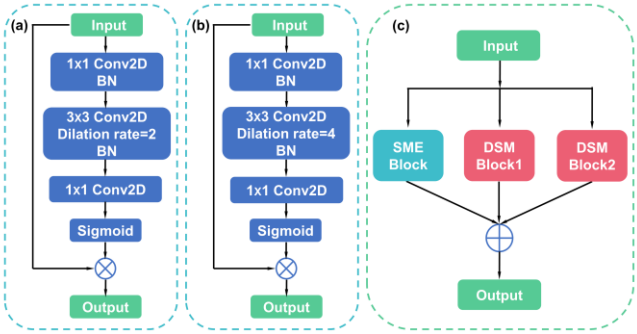

Figure 4. (a) Dilated spatial mapping (DSM) block 1, (b) DSM block 2, and (c) the entire DSMCA module.

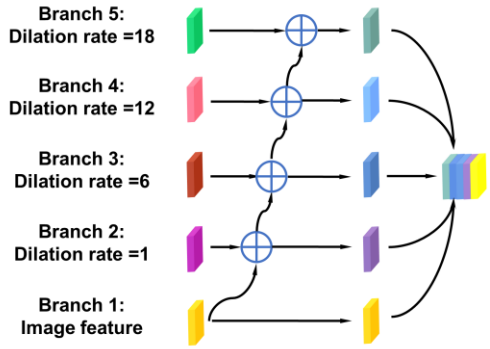

Figure 5. Branch layer fusion module

where X_{in} is the input to dilated spatial mapping and channel attention module. W_5, W_6, W_7, W_8, W_9, and W_{10} are the convolution layer parameters, $Dilation\_Conv(W_7 \cdot X_{in}, r = 2)$ and $Dilation\_Conv(W_9 \cdot X_{in}, r = 4)$ denote the results of performing dilation convolution (with a dilation rate of r) on $W_7 \cdot X_{in}$ and $W_7 \cdot X_{in}$, respectively. r is the dilation rate, and S is the output of the SME block. The terms of DSM_B_1 and DSM_B_2 denote the outputs of DSM block 1 and DSM block 2, respectively.

D. Branch layer fusion module

In order to improve the segmentation performance of lesions with different sizes, we have used five branches to capture multi-receptive field information. The image feature branch is utilized to extract global information, while the other four branches are used with different dilation rates (1, 6, 12, and 18) to capture contextual information at different scales. However, long-range information might be irrelevant due to the large dilation rate. To mitigate this problem, a BLF module is proposed to make a full use of information contained in these five branches, as shown in Fig. 5. All the features of five branches are integrated into an organized manner. The branches with the larger dilation rates are merged with the branches with

the smaller dilation rates in order to fuse multi-range context information and the fusion is realized by the 'add' operation.

E. Decoder

In the DeepLabv3+, a skip connection is included between the encoder and the decoder. Here, the DSMCA and MSFF modules are designed to aggregate multi-scale contextual

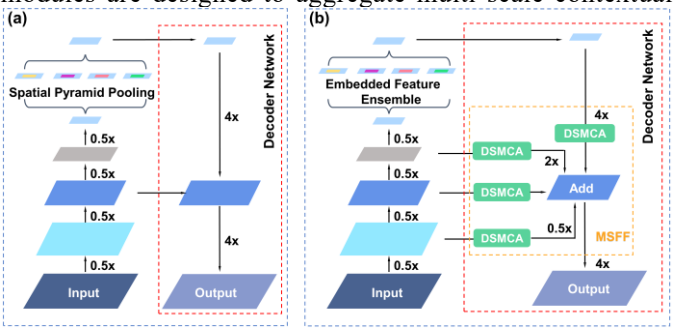

Figure 6. The encoder-decoder for (a) the DeepLabv3+ network and (b) our proposed model.

information between the encoder and the decoder (see Fig. 6). The output of BLF module is bilinearly up-sampled, whereas the low-level encoder feature maps are either up-sampled or down-sampled (depending on the feature map's resolution). The obtained features are then independently refined by the DSCMA module (corresponding to the channels in all four stages) and element-wise added. This constitutes the MSFF module, the output of which is then transmitted into two depth-wise separable convolution layers. A second bilinear up-sampling is then performed. Different from a simply single-stage fusion in the DeepLabv3+, the inputs of the MSFF module are derived from four different stages in the encoder network. We utilize DSMCA module in all the four stages to focus on more informative features.

F. Loss function

1) Preliminaries: Shapley values

The Shapley value is commonly used to demonstrate the contribution of each player to the total reward in a game [39]. Specifically, given a set of players $T=\{1,2,\cdots,t\}$, U denotes all the possible subsets of T. The number of subsets in the set T is 2^t. $g(S)$ denotes the reward obtained by the subset $S \subseteq T$ in the game. The Shapley value of player i in the game g, $\varphi(i|T)$, which represents the overall reward of i^{th} player, can be computed as follows.

$$\phi(i|T)=\sum_{s\subseteq T\setminus\{i\}} P_{Shapley}(S\,|\,T\setminus\{i\})\cdot[g(S\cup\{i\})-g(S)] \tag{10}$$

where $P_{Shapley}(S\,|\,H)=\dfrac{(|\,H\,|-|\,S\,|)!|\,S\,|!}{(|\,H\,|+1)!}$ is the probability of S being sampled, $S\subseteq H$.

2) Interactions between multiple tasks

In the game theory, the interactions among players may affect the final rewards [40]. Specifically, a model that contains multiple tasks could be considered as a game and the output of the model corresponds to the reward of the game. According to the game theory, two players may interact with each other to contribute to the reward of the game. Let us take an example of a model containing n tasks, and the task set is $T=\{1,2,\cdots,n\}$. Task i and task j in the set T have the interaction, which means that the results of task i and task j when they are done jointly are different with those when they are done individually. Based on this, we can define these two tasks as a special singleton task, e.g., $S_{ij}=\{i,j\}$. Thus, this multiple task method can be assumed to have $(n-1)$ tasks, e.g., $U'=U\setminus\{\{i\},\{j\}\}\cup S_{ij}$. Here the task i and task j are considered to be always absent or present simultaneously. In this way, the interaction $I_{(i,j)}$ between tasks i and j is defined as the contribution changes of S_{ij}, between the two cases, i.e., (1) when the tasks i and j are jointly performed, and (2) when tasks i and j are separately performed, as follows [36].

$$\begin{aligned} I_{(i,j)} &\overset{\text{def}}{=} \phi(S_{ij}\,|\,U')-[\phi(i\,|\,U\setminus\{j\})+\phi(j\,|\,U\setminus\{i\})] \\ &= \sum_{s\subseteq T\setminus\{i,j\}} P_{Shapley}(S\,|\,U\setminus\{i\})\cdot\Delta g(S,i,j) \end{aligned} \tag{11}$$

$$\Delta g(S,i,j)\overset{\text{def}}{=} g(S\cup\{i,j\})-g(S\cup\{j\})-g(S\cup\{i\})+g(S) \tag{12}$$

where $\phi(i\,|\,U\setminus\{j\})$ and $\phi(j\,|\,U\setminus\{i\})$ correspond to the contributions to the outputs of the model when the i^{th} and j^{th} tasks are done individually.

As explained previously, the proposed FCP-Net primarily perform two tasks, e.g., the segmentation task and classification task. These two tasks (or the relevant information) are considered as the input variables and form the player set. The predicted pixel-level probability map is considered as the reward. The output of the classification task is a probability value. The classification task is denoted with c and the segmentation task is denoted as s. M denotes all the possible subsets of $\{c,s\}$. Their baseline values are set as 0 [43]. When c is the only input, the output is too far away from the ground truth. In this condition, we set the reward of the game to be 0. Accordingly, the interactions between the segmentation and classification tasks can be calculated as follows.

$$\begin{aligned} I_{(c,s)} &= \sum_{D\subseteq M\setminus\{c,s\}} P_{Shapley}(D\,|\,M\setminus\{c,s\})\cdot\Delta g(D,c,s) \\ &= \frac{1}{12}\cdot[g(\{c,s\})-g(\{s\})] \end{aligned} \tag{13}$$

where $g(\{c,s\})$ denotes the results of the model that the segmentation and classification tasks are modeled jointly. $g(\{s\})$ denotes the result of the model that only contains the segmentation task. The absolute value of interaction $I_{(c,s)}$ represents the strength of the interaction. A small absolute value of $|I_{(c,s)}|$ indicates the small difference between the case when c and s are performed jointly and that when s was performed

individually. The detailed computation results of $I_{(c,s)}$ are provided in the APPENDIX.

3) Hybrid loss function

In the previously reported multi-task learning methods used in medical images, the loss function usually consists of the segmentation loss combined with the classification loss [34]. However, the interactions between segmentation and classification tasks have not been explored to improve the performance of the medical image segmentation. As we discussed above, the interaction $I_{(c,s)}$ may affect the results of the framework, and the absolute values of the $I_{(c,s)}$ demonstrate the strength of the interactions. Reducing the strength of the interactions between the segmentation task and classification task can make segmentation branch (s was performed individually) and interaction branch (c and s are performed jointly) collaboratively learn and teach each other throughout the training process. In this mutual learning process, these two branches effectively estimate the next most likely result. The segmentation branch and classification branch are optimized separately and then compared, and this process has been repeated until the best solution has been found. In this process, each branch finds the other's most probable results for training sample according to the peers. This will increase the posterior entropy of each branch, and help to converge on a more robust optimal solution with a better generalization performance.

In order to reduce the strength of interactions between these two tasks, in this paper, we design a new hybrid loss function which consists of three weighted functions.

The first term of the hybrid loss function is about the medical image segmentation task. The binary cross entropy loss function is selected to optimize the performance of segmentation task.

$$L_{seg} = -\frac{1}{N}\cdot\sum_{i=1}^{N}[q_i \cdot \log p_i + (1-q_i)\log(1-p_i)] \tag{14}$$

where N is the multiple of the height and the weight of the image. p_i denotes the prediction probability of i^{th} pixel, and q_i denotes the corresponding ground truth of i^{th} pixel.

The second term of the hybrid loss function is about the medical image classification task. The binary cross entropy loss function is selected to optimize the performance of classification task.

$$L_{cls} = -[y_i \cdot \log y_i' + (1-y_i)\log(1-y_i')] \tag{15}$$

where y_i denotes the label of the image and y_i' denotes the prediction probability of the image.

The third term of the hybrid loss function is about the interaction of the segmentation task and the classification task. To reduce the strength of the interaction and also improve the ability of generalization, we design an interaction loss function L_{int} as follows.

$$f(i,j) \overset{\text{def}}{=} 12 \cdot I_{(c,s)} = g(\{c,s\}) - g(\{s\}) \tag{16}$$

$$L_{int} = \frac{1}{W \times H}\cdot\sum_{i=0}^{W}\sum_{j=0}^{H}|\, f(i,j)\,| \tag{17}$$

where W and H denote weight and height of the model outputs. To simplify the calculation of L_{int}, we propose an interaction branch as shown in Fig. 1. The interaction branch produces the pixel-level probability maps when the segmentation task and classification task are cooperated with each other. We should address again that this only exists in the training stage.

The hybrid loss function L_{hybrid} is defined as follows.

$$L_{hybrd} = L_{seg} + \alpha \cdot L_{cls} + \beta \cdot L_{int} \tag{18}$$

where α, $\beta \in [0,1]$, in which α and β are hyper-parameters of the multi-task learning network.

IV. Experimental Results

A. Implementation details

1) Training settings

The FCP-Net was implemented in Keras using a Tensorflow backend on a GPU server with an Intel i5-7600K CPU running at 3.80 GHz, with 12 GB of RAM and an Nvidia GeForce 2080Ti GPU. All the training and testing steps were performed with the same hardware environment. Samples in these datasets were resized to 256×256 using a bilinear interpolation method. The RGB channels were rescaled to [0, 1] and maintained as the inputs to the proposed model. After this ablation study, the Adam optimizer was used for training, with an initial learning rate of 0.0001 and a batch size of 8. Label smoothening was applied to improve the results by reducing the overfitting. Training was ceased when the validation loss was remained a constant for 10 consecutive epochs. Fine-tuning models for these datasets were pre-trained using the PASCAL VOC dataset.

2) Evaluation metrics

Segmentation performance of the proposed model was assessed using the common evaluation metrics which were the same with those reported in Reference [41], including accuracy (ACC), sensitivity (SE), specificity (SP), dice coefficient (F1-Score), Jaccard index (JA), Dice, Intersection over Union (IoU), and the area under receiver operating characteristics curve (AUC).

B. Key component validation

1) Ablation Study

Ablation experiments were conducted to illustrate the effectiveness of our proposed FCP-Net. Here, we used the skin

TABLE I
A PERFORMANCE COMPARISON FOR DIFFERENT MODULES APPLIED TO THE ISIC-2017 DATASET

HLF	EFE	DSMCA1	BLF	MSFF	DSMCA2	F1-Score	SE	SP	ACC	JA	AUC
	✓	✓	✓	✓	✓	0.9035	0.8802	0.9832	0.9641	0.8282	0.9317
✓		✓	✓	✓	✓	0.9045	0.8855	0.9822	0.9633	0.8254	0.9336
✓	✓		✓	✓	✓	0.9056	0.8893	0.9817	0.9636	0.8271	0.9340
✓	✓	✓		✓	✓	0.9023	0.8791	0.9813	0.9625	0.8223	0.9313
✓	✓	✓	✓		✓	0.9019	0.8814	0.9818	0.9626	0.8213	0.9320
✓	✓	✓	✓	✓		0.9011	0.8817	0.9802	0.9619	0.8204	0.9324
✓	✓	✓	✓	✓	✓	0.9112	0.8912	0.9858	0.9669	0.8375	0.9371

HLF: hybrid loss function; EFE: embedded feature ensemble for separable convolution modules; DSMCA1: dilated spatial mapping and channel attention module used in the encoder; BLF: branch layer fusion modules; MSFF: multiple skip connections for multi-scale feature fusion; DSMCA2: dilated spatial mapping and channel attention module used in the decoder. Information fusion for five branches were conducted by using add and concatenate operations.

lesion segmentation as the example. In these experiments, training and testing of each network were performed with the same hardware environment using the ISIC-2017 training dataset. Trials with different modules were designed to illustrate the role played by each module in the network.

Table I compares the segmentation performance of six different FCP-Net modules. Fig. 7 provides a graphical visualization of the ablation experiment results. It is clear from the figure that the FCP-Net produces a much better performance of segmentation than those using the other algorithms, and also provides a much higher contrast in the regions of interest. We further conducted a t-test to understand the contribution of each component, and the obtained results are listed in Table II. All the proposed components (EFE, BLF, DSMCA1, DSMCA2, MSFF and HLF) have significant effects on the results obtained using this proposed method, especially for the MSFF and HLF modules.

After a comprehensive comparison, we observed that the proposed FCP-Net generally outperforms the other models and achieves a better performance, indicating that the proposed modules are effective for the segmentation task.

The EFE module encoded the multi-scale contextual information by combing depth-wise separable convolutions, SME-blocks, and residual connections. Residual connections then forcibly broke the symmetry of the neural network and reduced the degradation. The SME-block was a lightweight SE-block that included squeeze, excitation, and dropout operations. Global features at the channel level were then acquired via performing global averaging pooling (squeeze step). The excitation step captured the inter-channel dependencies for channel information in the squeeze step. Dropout layers were then added to avoid the model's overfitting.

TABLE II
P-VALUES FROM T-TEST ON DIFFERENT METRICS

Methods	F1-Score	SE	SP	ACC	JA	AUC
w/o HLF vs Ours	0.0002	0.001	0.009	0.001	0.0027	0.012
w/o EFE vs Ours	0.003	0.003	0.0003	0.001	0.005	0.029
w/o DSMCA1 vs Ours	0.005	0.052	0.0008	0.002	0.003	0.04
w/o BLF vs Ours	0.014	0.006	0.014	0.007	0.013	0.004
w/o MSFF vs Ours	0.002	0.016	0.004	0.001	0.002	0.012
w/o DSMCA2 vs Ours	0.002	0.012	0.007	0.001	0.003	0.022

It should be noted that although the SME block was originated from the SE block, these two blocks are quite different. Firstly, the SE block used the ReLU as the activation function after the first dense layer, but the ReLU did not perform very well for the high-level layers. We then replaced it with a Swish activation function. Secondly, there were too many channels in the input of the EFE module. SME block helped the network to be focused on important channels and avoid overfitting. In the SME block, we utilized the dropout to randomly discard some channels during the training stage, which can limit the interactions among different channels, making the information contained in the retained channels more conducive to the generalization of the model. In addition, we combined the SME block and residual block to mitigate the problems that the attention caused some high-frequency noises to propagate into deep layers of the network.

To demonstrate the effectiveness of dropout and residual block in the EFE module, we further performed the ablation studies using the ISIC 2017 training set. By discarding the

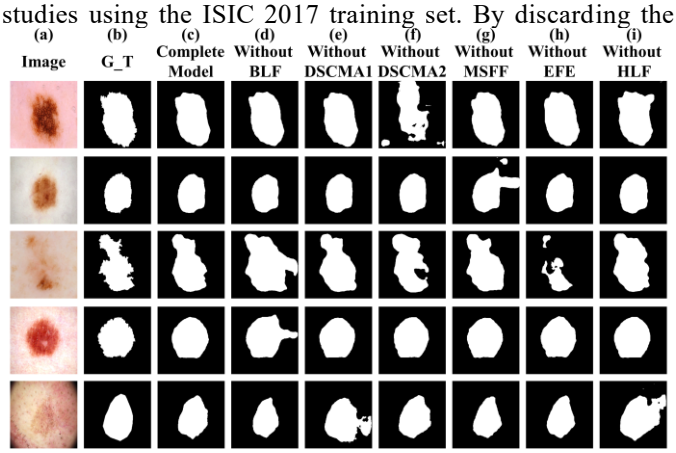

Figure 7. Results from the ablation experiment, including (a) the original dermoscopy image, (b) the ground truth, (c) results for FCP-Net, and results for models without (d) a BLF module, (e) a DSCMA1 module, (f) a DSMA2 module, (g) an MSFF module, and (h) an EFE module. (i) HLF.

dropout of SME block or residual blocks in the EFE module, we developed the FCP-Net without dropout or residual method, respectively. Training of the model was stopped when the validation loss was remained a constant for 10 consecutive epochs. The training stop times for the three cases of models without dropout, models without residual block, and FCP-Net are 42 epochs, 25 epochs, and 21 epochs, respectively. The results indicate that the dropout and residual block drastically decrease the training time. The metrics comparisons of these methods are listed in Table III. It can be seen that the dropout or residual method seriously affects the performance of the model, and our proposed EFE module shows a much better performance compared with the EFE modules without dropout or residual blocks.

To further verify the effectiveness of dropout layer in the SME block, we compared the mean standard deviation (MSD) of the attention weights by inputting the same targeted images with different resolutions (1.0×, 1.5× and 2.0×) to the FCP-Net and the FCP-Net without using a dropout method. We then collected the attention weights of SME block from five branches (containing EFE module) for these two cases. The

TABLE III
EFE MODULE ANALYSIS

Methods	F1-Score	SE	SP	ACC	JA	AUC
FCP-Net without dropout	0.8946	0.8625	0.9841	0.9602	0.8092	0.9232
FCP-Net without residual	0.8986	0.8784	0.9814	0.9613	0.8159	0.9298
Ours	0.9112	0.8912	0.9858	0.9669	0.8375	0.9371

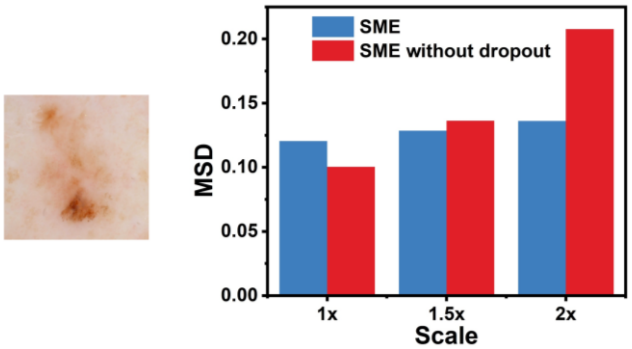

Figure 8. MSD of attention weights of SME block and SME block without dropout layer collected for the sample on 3 different resolutions (1.0× to 1.5× and 2.0×).

results shown in Fig. 8 clearly reveal that when the SME block has a dropout layer, the resolution change of the input has little effects on the MSD of the weights. This indicates that the EFE module ensures the stability of attention weight for targets with different scales.

The DSMCA consists of SME block and multiple dilated spatial mapping (DSM) blocks, which can recalibrate the feature maps separately along channel and space, and then combine the output. DSM blocks were used to explore the spatial relationships among different features, focusing on the regions of interest and integrating multi-scale contextual information. The above t-test results demonstrate that the proposed DSMCA has an important effect on the proposed method. To further validate the effectiveness of DSMCA, we incorporated the proposed DSMCA module into other classic methods such as U-Net and DeepLabv3+, and performed experiments on ISIC 2017. The results are listed in Table IV. It is evident that the proposed DSMCA module could also be utilized in other methods, suggesting that it is a crucial component for neural networks in medical applications.

The BLF module was used to extract and further merge the features from the five branches using adding and splicing operations. The MSFF module (containing several DSMCAs) was used to integrate the information from feature maps of different sizes. Up-sampling was applied to small feature maps, and down-sampling was applied to large feature maps. Information fusion was then applied to process the feature maps. The combination of EFE, DSMCA1, BLF, MSFF, and DSCMA2 modules provided a good solution to problems created by low grayscale variations and blurred boundaries, commonly observed in the medical image images.

2) Interaction Analysis

To explore the impact of the interactions between segmentation and classification tasks on the segmentation performance, we conducted experiments by setting different

TABLE IV
DSMCA MODULE ANALYSIS

Methods	F1-Score	SE	SP	ACC	JA	AUC
U-Net	0.7751	0.7127	0.9609	0.9023	0.6327	0.8368
DSMCA+U-Net	0.8138	0.7758	0.9707	0.9162	0.6861	0.8677
DeepLabv3+	0.79	0.7085	0.9737	0.9111	0.6529	0.8411
DSMCA+DeepLabv3+	0.8364	0.7718	0.9772	0.9287	0.7188	0.8745

TABLE V
HYPERPARAMETER EXPERIMENT RESULTS

ratio	F1-Score	SE	SP	ACC	JA	AUC
α=0, β=0	0.8169	0.8304	0.9507	0.9272	0.6904	0.8905
α=0, β=1	0.8859	0.9071	0.9658	0.9543	0.7952	0.9365
α=1, β=0	0.8453	0.7812	0.9837	0.9441	0.732	0.8724
α=1, β=1	0.8793	0.8299	0.986	0.9555	0.7847	0.9079
α=0.01, β=0.01	0.9098	0.9012	0.9857	0.9657	0.8345	0.9397
α=0.05, β=0.05	0.9061	0.8819	0.9843	0.9643	0.8283	0.9331
α=0.1, β=0.1	0.9112	0.8912	0.9858	0.9669	0.8375	0.9371

TABLE VI
HLF ANALYSIS

Methods	F1-Score	SE	SP	ACC	JA	Params	Inference Time/ms
U-Net	0.7751	0.7127	0.9609	0.9023	0.6327	31.03 M	15.8
FPM+U-Net	0.7896	0.6931	0.9651	0.9128	0.6523	31.15 M	18.3
HLF+U-Net (our)	0.8123	0.7614	0.9802	0.9171	0.6840	31.06 M	16.2
DeepLabv3+	0.79	0.7085	0.9737	0.9111	0.6529	41.25 M	16.5
FPM+DeepLabv3+	0.7888	0.6891	0.982	0.9129	0.6531	41.34 M	20.2
HLF+DeepLabv3+(our)	0.8063	0.7166	0.9843	0.9235	0.6715	41.26 M	17.6

weights of α and β in the HLF. We performed the experiments using the ISIC 2017 training dataset. The ISIC 2017 training set contains 2000 images, with 1250 samples used for training, 150 samples for validation, and 600 samples for testing.

As can be seen from Table V, our FCP-Net approach achieved the best performance when we set α=0.1 and β=0.1. This clearly indicates that the smaller values of α and β enabled the model being effectively trained and optimized to boost the performance of segmentation task. In addition, even if we did not consider the classification task in the loss function (e.g., α=0), the model still performed well. The possible reason is that the model would form pseudo-labels in an unsupervised manner.

To further validate that the proposed multi-task learning framework with the HLF is superior to the previous multi task learning framework, we utilized the HLF into other methods. We combined our HLF with the previously used architectures (e.g., U-Net and DeepLabv3+), and also combined previous multi-task learning framework which contains feature pass module (FPM) [19] with the previous architecture, in order to conduct experiments using the ISIC2017 dataset. The experiment results are summarized in Table VI. Compared with the previous multi-task learning framework which has an FPM to pass messages between segmentation and classification, our proposed HLF leads to less parameters, shorter inference time and better performance of segmentation.

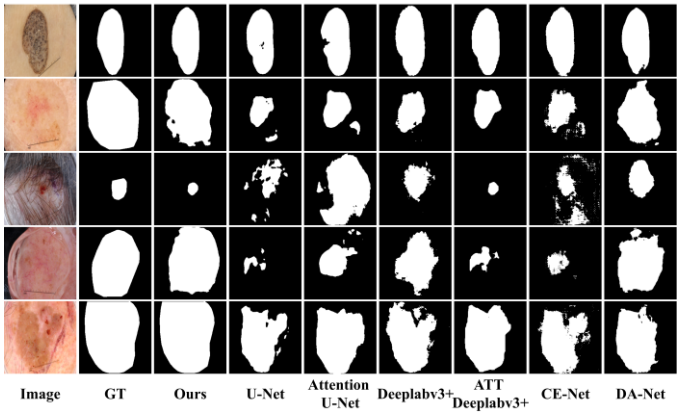

Figure 9. Segmentation results for different methods applied to the ISIC 2017 dataset.

TABLE VII
A PERFORMANCE COMPARISON FOR THE PROPOSED FCP-NET AND OTHER ALGORITHMS APPLIED TO ISIC 2017 DATASET

Methods	Years	F1-Score	SE	SP	ACC	JA
U-Net [1]	2015	0.7751	0.7127	0.9609	0.9023	0.6327
Attention-U-Net [2]	2018	0.7639	0.6504	0.9839	0.9051	0.6181
DeepLabv3+ [3]	2018	0.79	0.7085	0.9737	0.9111	0.6529
DA-Net [4]	2019	0.8123	0.7653	0.9633	0.9166	0.684
CE-Net [6]	2019	0.8082	0.7235	0.9794	0.9189	0.6782
KiU-Net [8]	2020	0.8063	0.7201	0.9731	0.9164	0.7464
DAGAN [11]	2020	0.8492	0.8351	0.9852	0.9354	0.7713
Att-DeepLabv3+ [9]	2020	0.8591	0.8054	0.9882	0.9414	0.8193
Ours	2021	0.9191	0.8799	0.989	0.963	0.8504

C. Skin Lesin Segmentation

Skin lesion segmentation plays a critical role in automatic and accurate diagnosis of melanomas [9], [42]. However, lesion segmentation is still a challenging task because of low contrast between lesions and normal skin areas.

The proposed FCP-Net was evaluated for skin lesion segmentation using the dataset of ISIC 2017, and ISIC 2018. The ISIC 2017 samples include the original images in a JPEG format and binary masks in a PNG format. The ISIC 2017 dataset includes samples of 2000 for training, 150 for validation, and 600 for testing. The height and width of ISIC 2017 images are ranged from 540 to 4499 and 722 to 6748, respectively. The ISIC 2018 dataset contains 2594 dermoscopy images. These images are derived from ISIC 2016 and ISIC 2017. However, the labels of classification tasks of these 2594 images were not supplied in the ISIC 2018. We obtained their labels from the ISIC 2016 and ISIC 2017, and also found that there were three same images which were labelled differently in the ISIC 2016 and ISIC 2017 (i.e., the annotated names of these image were 'ISIC_0000077', 'ISIC_0000511', and 'ISIC_0010094'). Therefore, we excluded these three images and used 2591 samples for training and testing for ISIC 2018. PH2 includes 200 samples in a BMP format (for both the original and mask images), consisting of 160 naevi and 40 melanomas with a fixed image size of 560×768.

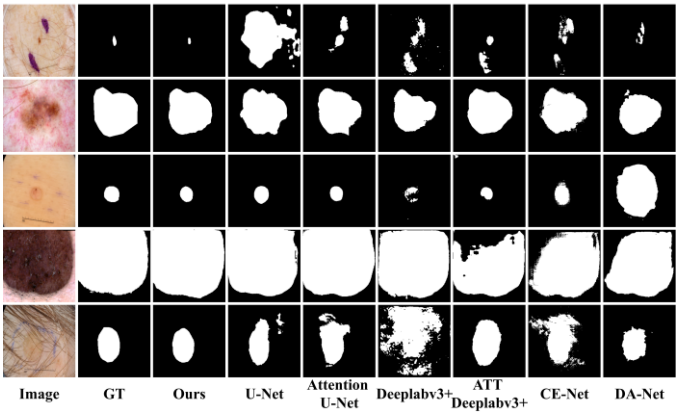

Figure 10. Segmentation results for different methods applied to the ISIC 2018 dataset.

TABLE VIII
A PERFORMANCE COMPARISON FOR THE FCP-NET AND OTHER ALGORITHMS APPLIED TO ISIC 2018 DATASET

Methods	Years	F1-Score	SE	SP	ACC	JA
U-Net [1]	2015	0.8393	0.8636	0.9478	0.9299	0.7231
Attention-U-Net [2]	2018	0.8861	0.8687	0.9753	0.9527	0.7955
DeepLabv3+ [3]	2018	0.8592	0.8721	0.9576	0.9395	0.7532
DA-Net [4]	2019	0.8706	0.8679	0.9662	0.9453	0.7708
CE-Net [6]	2019	0.8709	0.8798	0.9622	0.9448	0.7712
KiU-Net [8]	2020	0.8699	0.8692	0.9631	0.9364	0.7523
Att-DeepLabv3+ [9]	2020	0.8947	0.8689	0.9803	0.9567	0.8095
Ours	2021	0.924	0.907	0.985	0.968	0.8581

1) ISIC 2017

The proposed FCP-Net was evaluated using the ISIC2017 dataset. Sample segmentation results generated by the FCP-Net

are provided in Fig. 9. The results of the experiment are summarized in Table VII, and they were obtained from the proposed network and also from several state-of-the-art models. Comparing the evaluation metrics obtained from the proposed FCP-Net with those from the conventional methods, such as DeepLabv3+ and Att-DeepLabv3+ algorithms, we can conclude that the newly proposed modules improve the network's performance.

2) ISIC 2018

We further conducted experiments using the ISIC 2018 dataset to demonstrate the performance of the proposed model. This dataset contains 2591 valid samples. The original images and the corresponding ground truth annotations were applied. We utilized 1886 images for training, 141 for validation and 564 for testing. Fig. 10 shows the segmentation results obtained using the FCP-Net with the ISIC 2018 dataset. We can see that segmenting skin lesions using the FCP-Net method show much clearer boundary details. Table VIII summarizes the performance obtained using the proposed new method compared with those obtained using the other approaches. It can be seen that there are large differences among the results obtained from our proposed method and those from the other methods. In brief, the modules and new method we proposed can improve the performance of skin lesion segmentation.

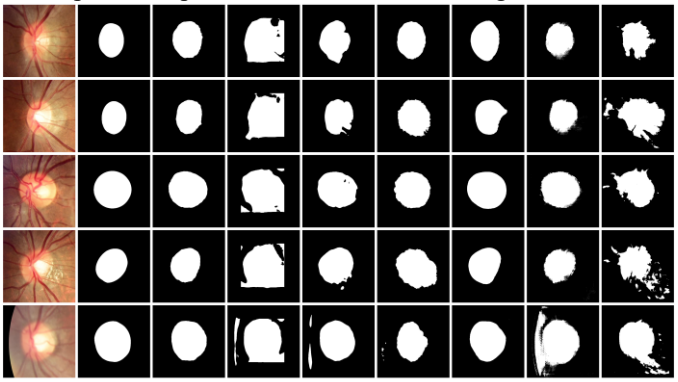

Figure 11. Segmentation results for different methods applied to the REFUGE dataset.

TABLE IX
SEGMENTATION PERFORMANCE OF THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE REFUGE DATASET.

Methods	Years	F1-Score	SE	SP	ACC	JA
U-Net [1]	2015	0.7146	0.9996	0.8454	0.8704	0.5559
Attention-U-Net [2]	2018	0.9033	0.9563	0.9688	0.9668	0.8237
DeepLabv3+ [3]	2018	0.8992	0.946	0.9694	0.9655	0.8168
DA-Net [4]	2019	0.7154	0.8721	0.8904	0.8874	0.5569
CE-Net [6]	2019	0.8875	0.9415	0.9651	0.9613	0.7978
Att-DeepLabv3+ [9]	2020	0.9134	0.9659	0.9711	0.9703	0.8407
Ours	2021	0.9183	0.9701	0.9724	0.9719	0.8488

D. Optic Disc Segmentation

Glaucoma is a disease that damages the optic nerves and often leads to irreversible vision loss, or even blindness. The optic disc (OD) segmentation is a crucial step for the diagnosis and analysis of glaucoma in the clinical applications [43]. Therefore, we evaluated the performance of the proposed framework on optic disc segmentation task. We tested the proposed method using the REFUGE dataset, which contains 400 training samples, 400 validation samples, and 400 testing samples. The image sizes of the training set, validation set, and test set are 2124×2056, 1634×1634, and 1634×1634, respectively. In the experiments, we extracted the regions of interest using the conventional digital image processing method and then resized the images into a size of 512×512. Fig. 11 shows examples of the visualization of OD segmentation results produced by the proposed model and other methods. Table IX compares the quantitative metrics obtained using our FCP-Net method and those from the other algorithms. All these results prove that our method has achieved a better performance compared with those obtained using the other state-of-the-art methods for the optic disc segmentation task, demonstrating that our method is a promising technique for the optic disc segmentation tasks.

E. Polyp Segmentation

Colorectal cancer is ranked as the third most common type of cancer [23]. Polyp segmentation from colonoscopy images is of great importance since it provides valuable information for diagnosis and surgery on the colorectal cancer. We evaluated the performance of the FCP-Net on polyp segmentation task.

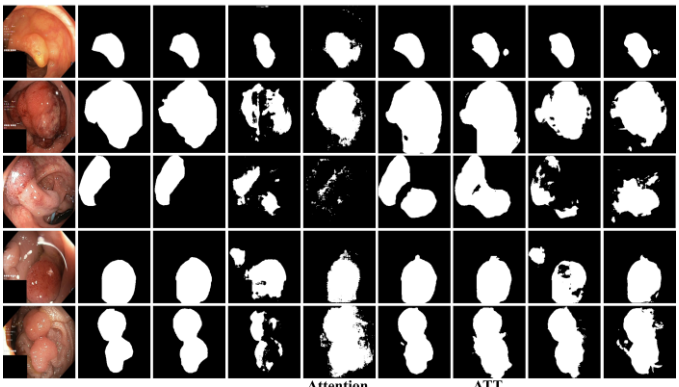

Figure 12. Segmentation results for different methods applied to the Kvasir-SEG dataset.

TABLE X
SEGMENTATION PERFORMANCE OF THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE KVASIR-SEG DATASET.

Methods	Years	SE	SP	ACC	IoU	Dice
U-Net [1]	2015	0.6088	0.9796	0.9187	0.5518	0.7112
Attention-U-Net [2]	2018	0.6822	0.9677	0.9208	0.5859	0.7389
DeepLabv3+ [3]	2018	0.8745	0.9726	0.9565	0.7675	0.8685
DA-Net [4]	2019	0.7429	0.9697	0.9324	0.6437	0.7832
CE-Net [6]	2019	0.7322	0.9771	0.9369	0.6559	0.7922
Att-DeepLabv3+ [9]	2020	0.8647	0.972	0.9579	0.7762	0.8739
Ours	2021	0.8669	0.9834	0.964	0.7977	0.8875

We tested the proposed method using the Kvasir-SEG dataset, which contains 1000 training samples (800 for training, 100 for validation, and 100 for testing). We followed the same settings and evaluation metrics as reported in Reference [44]. Fig. 12 shows examples of the visualization of polyp segmentation results produced by the proposed model and other methods. Table X compares the quantitative metrics obtained using the

proposed method and other algorithms, which clearly reveal that our method has achieved a better performance compared with those obtained using the other state-of-the-art methods for the polyp segmentation task. For example, compared with U-Net, the proposed FCP-Net achieves better performance by a large margin. Compared with the baseline DeepLabv3+, the proposed FCP-Net improves all six metrics and achieves better boundary information.

F. Breast ultrasound image segmentation

Breast cancer is a leading cause of death for women worldwide. Therefore, we applied the breast ultrasound image segmentation task as the demonstration of the fourth application. We evaluated our methods using the Breast Ultrasound (BUSI) dataset of 2020 [45]. The original BUSI dataset consists of 780 images (133 normal images, 437 benign images, and 210 malignant images) with an average image size of 500 × 500 pixels. The images are in the PNG format. The normal image has no lesion area, which means that the corresponding pixel-level label only contains one category. Based on the method reported in Reference [46], [47], we removed the normal images. The modified dataset consists of 647 samples. We randomly utilized 453 images for training, 65 for validation and 129 for testing. Fig. 13 shows examples of the visualization of polyp segmentation

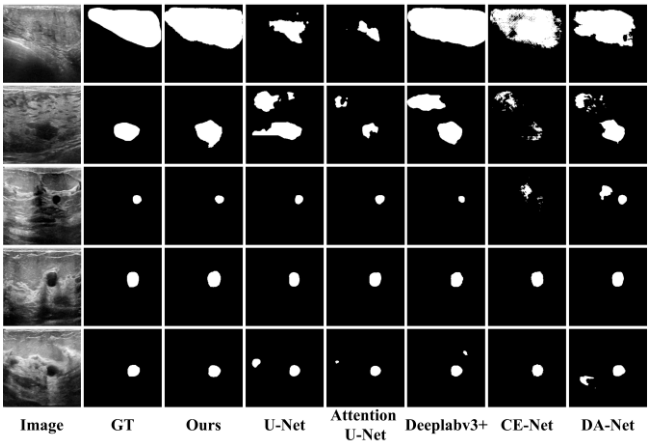

Figure 13. Segmentation results for different methods applied to the BUSI dataset.

TABLE XI
SEGMENTATION PERFORMANCE OF THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE BUSI DATASET.

Methods	Years	SP	ACC	AUC	Dice	IoU	HD
U-Net [1]	2015	0.9764	0.95	0.8193	0.6899	0.5265	6.324
Attention-U-Net [2]	2018	0.9928	0.9538	0.8299	0.6965	0.5886	6.304
DeepLabv3+ [3]	2018	0.9778	0.9631	0.8656	0.7848	0.6458	5.641
DA-Net [4]	2019	0.9793	0.959	0.8583	0.7509	0.6013	5.954
CE-Net [6]	2019	0.9828	0.9571	0.8291	0.7253	0.5689	5.989
Ours	2021	0.9875	0.967	0.869	0.7906	0.6537	5.331

results produced using our proposed model and the other methods. Table XI compares the quantitative metrics obtained using our proposed method and the other algorithms. Clearly our newly proposed FCP-Net archives a much better performance, especially for U-Net by a large margin (12.72 % for IoU, 10.07% for Dice, 4.97% for AUC). To reflect the boundary information, we obtained the boundary-based evaluation metrics of the Hausdorff distance. Compared with other methods, the proposed FCP-Net achieve a much better boundary information.

V. DISCUSSION

A. Framework discussion and network visualization

Delineating the lesions from their surrounding tissue is often a prerequisite step for diseases diagnosis and analysis. Efficiently training the deep neural networks for medical image segmentation is a challenging task due to the problems such as vanishing gradients and overfitting [48]. In this study, the FCP-Net was developed to improve automated lesion segmentation. The proposed model outperformed similar state-of-the-art methods when they were tested with several publicly available datasets.

In this study, we proposed and designed the EFE, DSMCA, BLF, and MSFF modules to effectively capture context information and fuse multi-scale features. Our proposed EFE module is quite different from the previous SE-block, which simply recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels. Our proposed EFE module adaptively recalibrates channel-wise feature responses and avoids the propagation of high-frequency noises

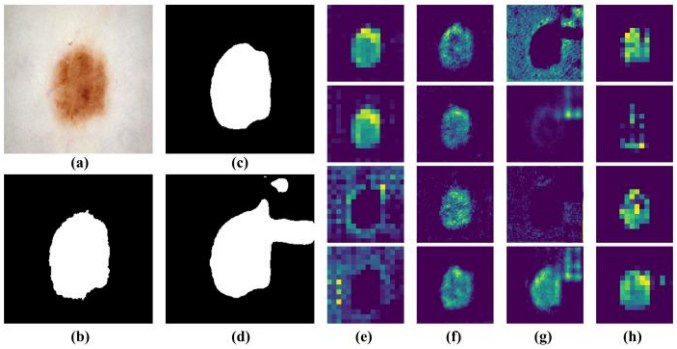

Figure 14. A visualization of method. (a) the original dermoscopy image, (b) the ground truth, (c) results of FCP-Net, (d) results of FCP-Net without MSFF module, (e), (f), and (h) feature maps in the end of BLF module, MSFF module, and Xception, respectively. (g) feature maps in FCP-Net without MSFF.

TABLE XII
A PERFORMANCE COMPARISON FOR THE FCP-NET AND BASELINES

Method	AUC			Parameters(M)
	ISIC2017	ISIC2018	PH2	
DeepLabv3+	0.8411	0.9148	0.9351	41
Att-DeepLabv3+	0.8468	0.9246	0.9561	57
Ours	0.9346	0.9459	0.9747	46

into much deeper layers by combining separable convolution, residual block and lightweight SME block. In addition, A BLF module was used in place of the direct branch splicing, which integrated context information contained in the five branches Furthermore, compared with the simply single-stage fusion in the DeepLabv3+, the proposed MSFF module can extract multi-scale features. In addition, to avoid redundant uses of low-level features, we proposed a DSMCA module, which combines multi-receptive field spatial attention mechanisms and channel

attention mechanisms to effectively capture multi-stage context information and enrich feature representation.

Last but not least, the previous methods commonly utilized feature pass modules or two-step learning to improve the performance of segmentation by learning shared information from classification. However, this will increase model parameters and complexity. Our proposed FCP-Net consists of segmentation, classification, and interaction branch. The classification branch and interaction branch were introduced to explore the interactions between medical image segmentation and classification task. The HLF was also proposed from a game theoretic view. As explained before, based on this HLF, we can make the segmentation, classification and interaction branches collaboratively learn and teach each other throughout the training process, thus taking full advantage of the conjoint information and improving the generalization performance. These training strategies significantly improve the segmentation results.

The FCP-Net improves the accuracy, F1-Score, sensitivity, specificity, and Jaccard similarity of automated lesion segmentation results, with only a small increase in the number of parameters. To demonstrate the good performance of FCP-Net, we compared the AUC and parameter values (shown in Table XII) using the proposed model and the baselines (DeepLabv3+). The number of FCP-Net parameters was ~80% of that of Att-DeepLabv3+ parameters [9]. The AUC values for the FCP-Net, applied to the ISIC2017, ISIC2018 and PH2 datasets, surpassed those of DeepLabv3+ by a significant margin.

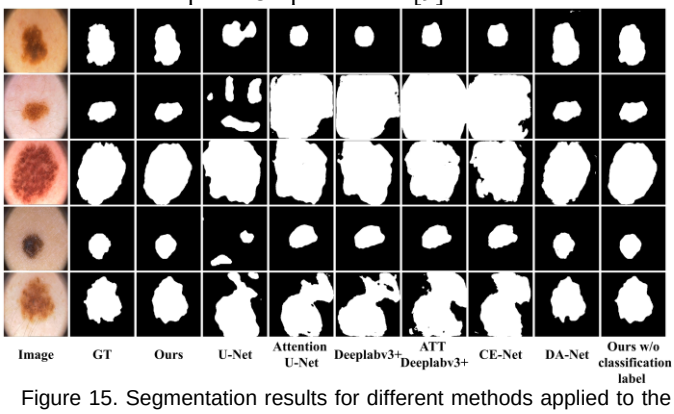

Figure 15. Segmentation results for different methods applied to the PH2 dataset.

TABLE XIII
SEGMENTATION PERFORMANCE OF THE PROPOSED FRAMEWORK AND OTHER METHODS ON THE PH2 DATASET.

Methods	F1-Score	SE	SP	ACC	JA	HD (mm)
U-Net [1]	0.9024	0.9091	0.9481	0.9352	0.8222	5.9895
Attention-U-Net [2]	0.9242	0.9169	0.9674	0.9509	0.8592	5.6425
DeepLabv3+ [3]	0.9145	0.9079	0.9623	0.9445	0.8425	5.1481
DA-Net [4]	0.9271	0.9035	0.9778	0.9535	0.8641	5.4566
CE-Net [6]	0.9429	0.9517	0.9675	0.9624	0.8921	5.1245
KiU-Net [8]	0.9407	0.9468	0.9612	0.9563	0.8721	5.1956
Att-DeepLabv3+ [9]	0.9396	0.9439	0.9683	0.9603	0.8861	5.1885
FCP+K-means	0.966	0.9787	0.9765	0.9775	0.9343	4.4273
Ours	0.9668	0.9642	0.9853	0.9784	0.9359	4.4171

Figure. 14 shows an example of feature extraction by the trained network. At the end of the extraction process, the network has learnt the detailed information such as contour boundaries, positions, and directions. This implies that the network has no need to crop images or remove information unrelated to lesion segments, such as the black circle on the periphery of the image. The boundaries in the feature maps (e.g., at the end of the BLF and MSFF modules) are similar, which suggests that our proposed model can be applied to more complex segmentation tasks. We have also demonstrated the effectiveness of MSFF module by visualizing some feature maps of FCP-Net and FCP-Net without MSFF, which can be revealed from the results shown in Fig. 14. We can observe that the results using the FCP-Net without the MSFF may be focused on the regions that do not contain any skin lesion, thus leading to the bad performance of segmentation.

B. The effectiveness of the framework under different conditions

Furthermore, we conducted practical applications on some segmentation tasks which did not have any image-level class labels. If the datasets of medical image segmentation do not have any image-level class labels, we can still use our method. In this case, the dataset is needed to encode with various methods, such as autoencoder network, K-means, and clustering based on a Gaussian mixture model. Experiments using the PH2 dataset were conducted to verify the effectiveness of the FCP-Net. The PH2 datasets are a set of dermoscopic images widely applied as a benchmark for algorithm validation, and they contain both the ground truth in pixel-level label and image-level class label. Other experiments using the modified PH2 dataset (e.g., the image-level class labels were artificially removed before the experiments) were also conducted. For the experiments using the modified PH2 dataset, we used the K-means to annotate the image-level class label, and then conducted the experiments. To reflect the boundary information, we also increased the boundary-based evaluation metrics of the Hausdorff distance [49].

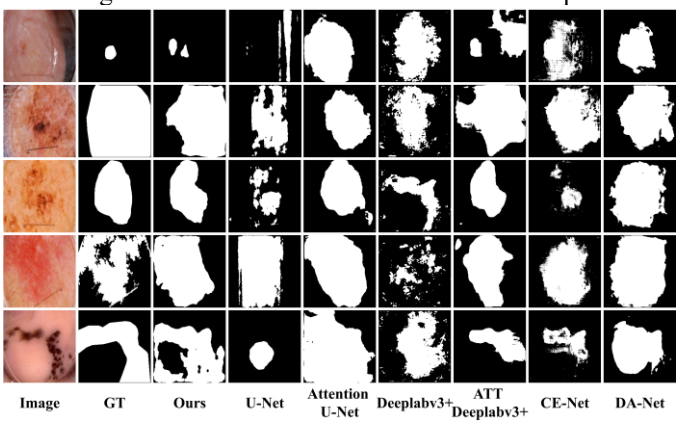

Figure 16. Cases where our results are not good enough

The obtained results are presented in Fig. 15 and the data are listed in Table XIII. Compared with the experimental results obtained from the original PH2 dataset, those from the modified PH2 dataset (e.g., the image-level class label was artificially removed) using our proposed method are quite similar. Furthermore, compared with those obtained using the other

state-of-the-art methods, our results of experiments using the modified PH2 dataset are more suitable. These experiment results prove that our proposed method is quite effective when there are no classification data available.

Although the proposed FCP-Net has achieved promising performance on those datasets, our method has some limitations. Similar to those of most previously used methods, our method may suffer from problems of segmenting when the images have very low contrast, as shown in Fig. 16. However, we still achieved better performance compared with other state-of-the-art methods. In addition, as can be seen from Table XII, the proposed method still has many parameters, which means that the FCP-Net requires excellent hardware conditions and environment. Furthermore, to extract features in a light weight manner, we expanded the receptive field by using multiple dilation convolution layers, which may lead to inefficiently modeling of long-range feature dependencies and cause non-optimal discriminative feature representations associated with each semantic class.

VI. Conclusion

In this study, we proposed an FCP-Net with a hybrid loss function in order to effectively capture the global and multi-scale context information and explore the interactions between medical image segmentation and classification, thus enhancing the performance of segmentation. We have designed EFE, DSMCA, BLF, and MSFF modules to effectively capture context information and fuse multi-scale features. The EFE module was proposed to adaptively recalibrate the attention weights and explore the complementariness between attention mechanisms and residual blocks. These overcame the problems that the attention weights may be unstable when inputting the same targeted images with different resolutions, and the attention may cause some high-frequency noises to propagate into deep layers of the network. The DSMCA, BLF, and MSFF modules were proposed to efficiently identify spatial correlations between integrate multi-receptive field features from different branches, and establish multiple skip connections for the enhanced fusing features. In addition, the classification branch and interaction branch were introduced to explore the interactions between medical image segmentation and classification. An HLF was proposed from a game theoretic view. Based on this HLF we can make the segmentation, classification and interaction branch collaboratively learn and teach each other throughout the training process, thus fully taking advantage of the conjoint information and improving the generalization performance. Experiment results, including skin lesion segmentation, optic disc segmentation, polyp segmentation, and breast ultrasound image segmentation, showed that the proposed FCP-Net is a promising method for medical image segmentation tasks.

Appendix

This section presents the detailed methods for calculating, $I_{(c,s)}$, which represents the interactions between the classification task (c) and the segmentation task (s). M denotes all possible subsets of $\{c,s\}$. The interactions between classification task and segmentation task are calculated as follows.

$$
\begin{aligned}
&I_{(c,s)} \\
&= \sum_{D \subseteq M \setminus \{c,s\}} P_{Shapley}(D \mid M \setminus \{c,s\}) \cdot \Delta g(D,c,s) \\
&= \frac{2!}{4!} \cdot [(g(\{\emptyset\} \cup \{c,s\})) - g(\{\emptyset\} \cup \{s\}) - g(\{\emptyset\} \cup \{c\}) + g(\{\emptyset\})] \\
&+ \frac{2!}{4!} \cdot [(g(\{c\} \cup \{c,s\})) - g(\{c\} \cup \{s\}) - g(\{c\} \cup \{c\}) + g(\{c\})] \\
&+ \frac{2!}{4!} \cdot [(g(\{s\} \cup \{c,s\})) - g(\{s\} \cup \{s\}) - g(\{s\} \cup \{c\}) + g(\{s\})] \\
&= \frac{1}{12} \cdot [g(\{c,s\}) - g(\{c\} \cup \{c\}) - g(\{s\} \cup \{s\})] \\
&= \frac{1}{12} \cdot [g(\{c,s\}) - g(\{c\}) - g(\{s\})] \\
&= \frac{1}{12} \cdot [g(\{c,s\}) - g(\{s\})]
\end{aligned}
$$

where $D \subseteq M \setminus \{c,s\}$ equivalent to $D = \{\{\emptyset\},\{c\},\{s\}\}$. $g(\{\emptyset\}) = g(\{c\}) = 0$.

Acknowledgment

This work was supported by the General Program of National Natural Science Foundation of China (NSFC No. 52075162), Innovation Leading Program of New and High-tech Industry of Hunan Province (2020GK2015), the Natural Science Foundation of Hunan Province (2021jj20018), the Natural Science Foundation of Changsha (kq2007026), the Key Research Project of Guangdong Province (2020B0101040002), and International Exchange Grant (IEC/NSFC/201078) through the Royal Society, UK and the NSFC.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234-241.

[2] O. Oktay *et al.* (2018). "Attention U-Net: Learning where to look for the pancreas,". [Online]. Available: https://arxiv.org/abs/*1804.03999*.

[3] 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.

[4] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 3146-3154.

[5] A. Saez, C. Serrano, and B. Acha, "Model-based classification methods of global patterns in dermoscopic images," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1137-1147, 2014.

[6] Z. Gu *et al.*, "Ce-net: Context encoder network for 2d medical image segmentation," *IEEE Trans. Med. Imag.,* vol. 38, no. 10, pp. 2281-2292, 2019.

[7] S. Pathan, K. G. Prabhu, and P. C. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions—A review," *Biomed. Signal. Process. Control.,* vol. 39, pp. 237-262, 2018.

[8] J. M. J. Valanarasu, V. A. Sindagi, I. Hacihaliloglu, and V. M. Patel, "KiU-Net: Towards accurate segmentation of biomedical images using over-complete representations," in *Proc. MICCAI*, 2020, pp. 363-373.

[9] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, and S. Escalera, "Attention DeepLabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 251-266.

[10] Y. LeCun, Y. Bengio, and G. J. n. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015.

[11] B. Lei et al., "Skin lesion segmentation via generative adversarial networks with dual discriminators," *Med. Image Anal.*, vol. 64, 2020.

[12] G. Wang, P. Luo, L. Lin, and X. Wang, "Learning object interactions and descriptions for semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 5859-5867.

[13] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.,* vol. 39, no. 10, pp. 3008-3018, 2020.

[14] Y. Xue, T. Xu, H. Zhang, L. Long, and X. Huang, "Segan: Adversarial network with multi-scale l1 loss for medical image segmentation," *Neuroinformatics,* vol. 16, no. 3, pp. 383-392, 2018.

[15] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3156-3164.

[16] H. Wu, J. Pan, Z. Li, Z. Wen, and J. Qin, "Automated skin lesion segmentation via an adaptive dual attention module," *IEEE Trans. Med. Imag.,* vol. 40, no. 1, pp. 357-370, 2020.

[17] 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.

[18] M. Sarker *et al.*, "SLSNet: Skin lesion segmentation using a lightweight generative adversarial network," *Expert Syst. Appl.,* p. 115433, 2021.

[19] S. Chen, Z. Wang, J. Shi, B. Liu, and N. Yu, "A multi-task framework with feature passing module for skin lesion classification and segmentation," in *Proc IEEE 15th Int. Symp. Biomed. Imaging. (ISBI)*, 2018, pp. 1126-1129.

[20] T. He, J. Hu, Y. Song, J. Guo, and Z. Yi, "Multi-task learning for the segmentation of organs at risk with label dependence," *Med. Image Anal.*, vol. 61, p. 101666, 2020.

[21] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 7132-7141.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3431-3440.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770-778.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. Conf. AAAI. Artif. Intell.*, 2017.

[25] Y. Qin *et al.*, "Autofocus layer for semantic segmentation," *Proc. MICCAI*, 2018, pp. 603-611.

[26] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2881-2890.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. Yuille, and m. intelligence, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern. Anal. Mach. Intell.,* vol. 40, no. 4, pp. 834-848, 2017.

[28] M. Sarker et al., "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Proc. MICCAI*, 2018, pp. 21-29.

[29] Y. Xue, T. Xu, and X. Huang, "Adversarial learning with multi-scale loss for skin lesion segmentation," in *Proc IEEE 15th Int. Symp. Biomed. Imaging. (ISBI)*, 2018, pp. 859-863.

[30] M. Prathiba, D. Jose, and R. Saranya, "Automated melanoma recognition in dermoscopy images via very deep residual networks," in *IOP Conf. Ser., Mater. Sci. Eng.,*, 2019, vol. 561, no. 012107.

[31] L. Yu, H. Chen, Q. Dou, J. Qin, and P. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.,* vol. 36, no. 4, pp. 994-1004, 2016.

[32] I. Gonzalez-Diaz and h. informatics, "Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE J. Biomed. Health Inform.,* vol. 23, no. 2, pp. 547-559, 2018.

[33] S. Hong, H. Noh, and B. J. a. p. a. Han, "Decoupled deep neural network for semi-supervised semantic segmentation," in *Proc. NIPS,* 2015.

[34] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Trans. Med. Imag.,* vol. 39, no. 7, pp. 2482-2493, 2020.

[35] S. Vandenhende, S. Georgoulis, and L. Van Gool, "Mti-net: Multi-scale task interaction networks for multi-task learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 527-543.

[36] H. Zhang, S. Li, Y. Ma, M. Li, Y. Xie, and Q. Zhang. (2020). "Interpreting and boosting dropout from a game-theoretic view," . [Online]. Available: https://arxiv.org/abs/*1705.04977*.

[37] L. Kaiser, A. N. Gomez, and F. Chollet. (2017). "Depthwise separable convolutions for neural machine translation,". [Online]. Available: https://arxiv.org/abs/*1706.03059*.

[38] P. Ramachandran, B. Zoph, and Q. Le. (2017). "Searching for activation functions," . [Online]. Available: https://arxiv.org/abs/*1710.05941*.

[39] L. J. A. M. S. Shapley, Contributions to the Theory of Games, ed. by HW Kuhn, and A. Tucker, "A value fo n-person Games," *Ann. Math. Study28, Contributions to the Theory of Games,* pp. 307-317, 1953.

[40] M. Grabisch and M. J. I. J. o. g. t. Roubens, "An axiomatic approach to the concept of interaction among players in cooperative games," *International Journal of game theory,* vol. 28, no. 4, pp. 547-565, 1999.

[41] N. Codella *et al.* (2019). "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic),". [Online]. Available: https://arxiv.org/abs/*1902.03368*.

[42] H.-C. Shin et al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285-1298, 2016.

[43] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. J. I. t. o. m. i. Heng, "Patch-based output space adversarial learning for joint optic disc and cup segmentation," *IEEE Trans. Med. Imag.,* vol. 38, no. 11, pp. 2485-2495, 2019.

[44] D.-P. Fan *et al.*, "Pranet: Parallel reverse attention network for polyp segmentation," in *Proc. MICCAI*, 2020, pp. 263-273: Springer.

[45] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief,* vol. 28, p. 104863, 2020.

[46] H. Zou, X. Gong, J. Luo, T. J. C. M. Li, and P. i. Biomedicine, "A Robust Breast ultrasound segmentation method under noisy annotations," *Comput. Methods. Programs. Biomed. ,* vol. 209, p. 106327, 2021.

[47] M. Byra *et al.*, "Breast mass segmentation in ultrasound with selective kernel U-Net convolutional neural network," *Biomed. Signal. Process. Control.,* vol. 61, p. 102027, 2020.

[48] Y. Yuan, M. Chao, and Y. C. Lo, "Automatic Skin Lesion Segmentation Using Deep Fully Convolutional Networks With Jaccard Distance," *IEEE Trans. Med. Imag.*, vol. 36, no. 9, pp. 1876-1886, Sep. 2017.

[49] D. P. Huttenlocher, G. A. Klanderman, W. Rucklidge, and m. intelligence, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850-863, 1993.