

Boundary-aware Context Neural Network for Medical Image Segmentation

Ruxin Wang, Shuyuan Chen, Chaojie Ji, Jianping Fan and Ye Li, *Senior Member, IEEE*

Abstract—Medical image segmentation can provide a reliable basis for further clinical analysis and disease diagnosis. The performance of medical image segmentation has been significantly advanced with the convolutional neural networks (CNNs). However, most existing CNNs-based methods often produce unsatisfactory segmentation mask without accurate object boundaries. This is caused by the limited context information and inadequate discriminative feature maps after consecutive pooling and convolution operations. In that the medical image is characterized by the high intra-class variation, inter-class indistinction and noise, extracting powerful context and aggregating discriminative features for fine-grained segmentation are still challenging today. In this paper, we formulate a boundary-aware context neural network (BA-Net) for 2D medical image segmentation to capture richer context and preserve fine spatial information. BA-Net adopts encoder-decoder architecture. In each stage of encoder network, pyramid edge extraction module is proposed for obtaining edge information with multiple granularities firstly. Then we design a mini multi-task learning module for jointly learning to segment object masks and detect lesion boundaries. In particular, a new interactive attention is proposed to bridge two tasks for achieving information complementarity between different tasks, which effectively leverages the boundary information for offering a strong cue to better segmentation prediction. At last, a cross feature fusion module aims to selectively aggregate multi-level features from the whole encoder network. By cascaded three modules, richer context and fine-grain features of each stage are encoded. Extensive experiments on five datasets show that the proposed BA-Net outperforms state-of-the-art approaches.

Index Terms—Convolutional neural network, deep learning, medical image segmentation.

I. INTRODUCTION

IMAGE segmentation plays an important role in medical image analysis, which aims to address pixel-wise and fine-grained lesion recognition [1], [2]. With the development and

popularization of medical imaging technology and equipment, ultrasound, magnetic resonance imaging (MRI), computed tomography (CT) and other imaging modalities provide an intuitive and effective way to diagnose and scan different kinds of diseases. These techniques have been widely used in daily clinical research and treatment planning. For different types of clinical applications, segmentation has been adopted as a key step of image analysis, such as lung segmentation in CT images [3], skin lesion segmentation in dermoscopy images [4], colorectal cancer segmentation in endoscopy images [5] and cell segmentation in microscopy images [6]. Accurate lesion detection is critical to provide a reliable basis for further clinical analysis [7], disease diagnosis [8], therapy planning [9] and prognosis evaluation [10]. High precision is typically required in lesion segmentation which need to segment the focused parts and extract relevant features accurately [11].

With the increasing number of medical images and the development of Artificial Intelligence (AI), computer-assisted diagnosis technology can effectively assist professional clinicians to improve the accuracy and efficiency of analysis. However, automatic lesion (organ or tissue) recognition in medical image remains a complex and challenging task [12], [13]. At first, lesion regions have various sizes and shapes for different individuals. For some diseases, obvious individual differences increase the difficulty of recognition. Fig. 1 shows two examples for skin lesion and colorectal polyp. Secondly, The low contrast between the lesions and background also brings great challenge to segmentation. In particular, the focused area usually contains complex tissues and organs, which makes it very difficult to distinguish these confusing boundary pixels. In addition, some artifacts and imaging noise also impede the accuracy of segmentation.

In the past few decades, a large group of automatic analysis algorithms for medical image segmentation have been proposed [14]–[17], which can be roughly classified into three categories: gray level-based [18], texture-based [19] and atlas-based methods [20]. Although these approaches have enhanced the performance of automated segmentation through extracting different kinds of pixel and region features, they still have some common defects: 1) Traditional methods often design low-level hand-crafted features and make heuristic hypotheses, which usually restricts the performance of prediction with complex scenarios. Moreover, abundant available information in the original image is neglected. 2) The robustness to artifacts, image quality and intensity inhomogeneity is low, which heavily depends on effective pre-processing.

Recently, due to the remarkable successes of deep learning in computer vision, deep convolutional neural network (CNNs)

This work was funded by National Natural Science Foundation of China (U1913210, 61771465), Major Special Project of Guangdong Province (2017B030308007), Strategic Priority CAS Project XDB38000000, Shenzhen Basic Research Projects (JCYJ20180703145202065), Shenzhen Science and Technology Innovation Project (JSGG20170823144843046).

Ye Li and Jianping Fan are the *Corresponding author*.

Ruxin Wang is with Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (e-mail: rx.wang@siat.ac.cn)

Shuyuan Chen is with the School of Software Engineering, University of Science and Technology of China, Hefei, and Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (e-mail: sa517020@mail.ustc.edu.cn)

Chaojie Ji is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (e-mail: cj.ji@siat.ac.cn)

Jianping Fan is with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (e-mail: jp.fan@siat.ac.cn)

Ye Li is with Joint Engineering Research Center for Health Big Data Intelligent Analysis Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. (e-mail: ye.li@siat.ac.cn)

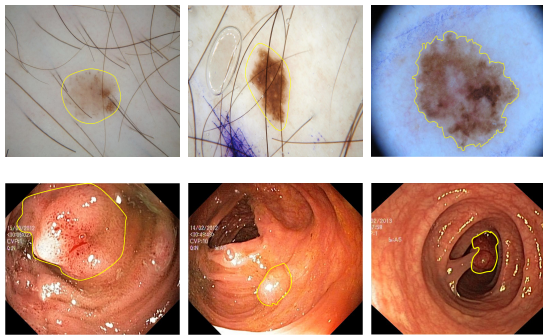


Fig. 1. Examples of two representative medical images. The first row shows the skin lesion in the dermoscopy image and the second row indicates the colorectal polyp in endoscopy images. In each image, yellow solid line refers to the target boundary.

has emerged as a promising alternative for medical image segmentation [21]–[23], which successfully overcomes the limits of traditional hand-crafted features. Most state-of-the-art medical image segmentation approaches are based on encoder-decoder network architecture, among which the most representative methods are U-net [24] and fully convolutional network (FCN) [25]. The network framework is designed in an end-to-end way with pixel-wise supervision. In encoder stage, input image obtains high-level semantic feature representations through consecutive convolution operations. Then the top features of encoder network are employed to generate the predicted segmentation mask by progressive upsampling (uppooling or deconvolution) method in decoder network. Although convolutional neural networks have shown their advantages in the medical image segmentation task, most existing CNNs-based methods still suffer from inaccurate object boundaries and unsatisfactory segmentation results. This is caused by the limited context information and inadequate discriminative feature maps after consecutive pooling and convolution operations. In order to accurately recognize object, it is necessary to extract and aggregate high-level semantic features with low-level fine details simultaneously. Overall, how to learn richer context is still a challenge of segmentation algorithm for improving the recognition performance.

Inspired by above analysis, we propose a novel convolutional network framework based on boundary-aware (BA-Net) for medical image segmentation, which follows the classical encoder-decoder structure. Specifically, in each stage of encoder network, pyramid edge feature extraction module (PEE) is proposed for obtaining edge information with multiple granularities firstly. Object boundaries define the shape of the object, and thus provide complementary cues for segmenting target objects. For getting richer knowledge about the sample, in each stage of encoder network, we design a mini multi-task learning module (mini-MTL) and jointly supervise segmentation and boundary map prediction during training. Furthermore, to take full advantage of features from different task, an interactive attention (IA) is proposed. IA makes use of the interactive information from different tasks to supervise the modeling of the target area recognition which is helpful to refine the segmentation performance. At last, a cross feature

fusion module (CFF) is presented by selectively aggregate multi-level features from the whole encoder network, which further captures valuable context and preserves fine spatial information. By cascaded three modules, richer context and fine-grain features of each stage are encoded. In decoder network, we integrate these feature maps to get the final segmentation prediction sequentially. Finally, we evaluate our BA-Net on the multiple public medical image datasets and achieve consistent performance improvements on them.

In summary, the contributions of this work are four-fold:

- 1) We put forward a novel boundary-aware context neural network for 2D medical image segmentation, which employs PEE, mini-MTL and CFF modules to produce richer contextual information for guiding the decoder processing.
- 2) We design the PEE module and mini-MTL module with embedded interactive attention for fully mining the contextual features at the same level, which effectively leverage the boundary information for offering a strong cue to better segmentation prediction.
- 3) We build a CFF module to selectively incorporate cross-level features from other stage of the encoder network into current stage. In this way, information complementation among different levels of features is realized.
- 4) We conduct comprehensive experiments and achieve outstanding state-of-the-art segmentation performance for different tasks including skin lesion segmentation, colorectal polyp segmentation, lung segmentation and optic disc segmentation. The experimental results convince the efficiency of the proposed method.

The remainder of this paper is organized as follows: Section II reviews the recent development of medical image segmentation. Section III presents our proposed segmentation neural network in detail. In Section IV, we evaluate the proposed model, and compare it with the state-of-the-art methods on the multiple public datasets. And related analysis of our method is discussed. Finally, Section V concludes this paper and prospects some future works.

II. RELATED WORK

In this section, related conventional and CNN-based segmentation methods for medical image segmentation are briefly reviewed.

A. Conventional Segmentation Methods

Over the past decades, various models have been introduced including gray level-based, texture-based and atlas-based models. Early methods based on gray level features mainly contain histogram statistic, edge detection and region growing strategies. For example, Carballido-Gamio et al. [26] applied the Normalized Cuts with local histograms of brightness to segment vertebral bodies from sagittal T1-weighted magnetic resonance images of the spine. Chung et al. [27] presented a partial-differential equations-based framework for detecting the boundary of skin lesions in dermoscopy images, which the object area is segmented either by the geodesic active contours or the geodesic edge tracing model. Nguyen et al. [28] proposed the watersnakes model for image

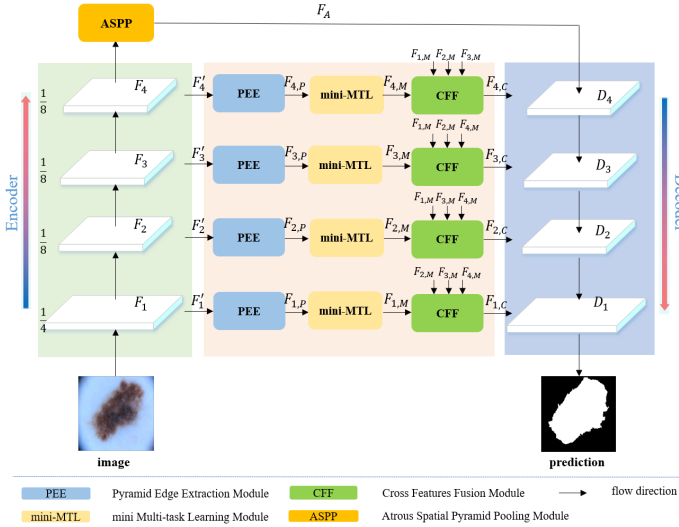


Fig. 2. The framework of the proposed model. In each stage of encoder network, we firstly obtain multiple granularities edge features by pyramid edge extraction module (PEE). Then the mini multi-task learning module (mini-MTL) jointly learns to segment object masks and detect lesion boundaries. At last, multi-level features from the whole encoder network are selectively aggregated by cross feature fusion module (CFF), which further captures valuable context and preserves fine spatial information.

segmentation by adding the contour length to the energy function. It integrates the strengths of watershed segmentation and energy based segmentation. Xie et al. [29] presented a novel texture and shape priors extracted by applying a bank of Gabor filters for kidney segmentation in ultrasound images. In Bazin’s work [30], they designed a segmentation framework based on both topological information and statistical atlases of brain anatomy which constrained topological equivalence between the prediction and the atlas. Although these models such as thresholding and region growing approaches are able to implement, but the performance is restricted in that the selection of threshold and region division criteria is greatly affected by the image intensity or texture information. Also hand-craft features for the segmentation are heavily relied on the experience of the researchers.

B. CNN-based Segmentation Frameworks

In recent years, deep convolutional neural networks have been successfully applied to a wide variety of problems in computer vision, which shows striking profit using encoder-decoder framework for image segmentation tasks [31]–[33]. In encoder network, image content is encoded by multiple convolutional layers from low-level to high-level. And in decoder part, the prediction mask is obtained by multiple upsampling (uppooling or deconvolutional) layers. In particular, image feature representation and context extraction are very crucial for segmentation task. For example, Chen et al. [34] proposed a segmentation framework named DeepLab which tailored an atrous spatial pyramid pooling module to encode the multi-scale contextual information by parallel dilated convolution. Zhao et al. [35] employed a pyramid pooling module with multiple pooling scales to capture multi-scale context in encoder network. Li et al. [36] presented a new dense deconvolutional

network for skin lesion segmentation based on residual learning. It captures fine-grained multi-scale features of image for segmentation task by dense deconvolutional layers, chained residual pooling and auxiliary supervision. Xue et al. [37] proposed a end-to-end adversarial critic neural network with a multi-scale L1 loss function for medical image segmentation, which forces to learn both global and local features. Zhou et al. [38] tailored a collaborative network architecture to jointly improve the performance of disease classification and lesion segmentation by semi-supervised learning with attention mechanism. Chen et al. [39] proposed an unsupervised domain adaptation method to effectively tackle the problem of domain shift and achieved cross-modality image segmentation. Overall, effective extraction of image context is important for improving segmentation performance.

III. METHODOLOGY

In this section, we describe the construction of the proposed boundary-aware context neural network and the design methods of the three core modules (i.e. PEE, mini-MTL and CFF). Details of the proposed method are introduced as follows.

A. Overview

As shown in Fig. 2, the proposed BA-Net has an encoder-decoder architecture design and starts with ResNet [40] as backbone (pre-trained on ImageNet [41]). In encoder network, the last global pooling and fully-connected layers of ResNet are removed firstly which only retains one convolution and four residual convolution blocks for primary features extraction. Without loss of generality, for an input image, we denote the output features of four residual blocks as $F_i, i \in \{1, 2, 3, 4\}$. To further enlarge the receptive fields, the last two blocks in ResNet are modified using atrous convolution (atrous rate = 2) and maintain the same spatial resolution as the previous block by removing the pooling operation in our work. Thus, the output sizes of each block are 1/4, 1/8, 1/8 and 1/8 of the input image. In addition, an atrous spatial pyramid pooling (ASPP) module [34] is employed on the top feature maps of last residual block for capturing and encoding multi-scale features. The ASPP module comprises of four parallel atrous convolutions with different atrous rates and one global average pooling. The output features of ASPP are concatenated by upsampling and one 1×1 convolution (with 256 channels), which further integrates and compresses the feature maps. In order to produce richer contextual information for guiding the decoder processing, we tailor three modules to fully mine the features in same level and aggregate other features from different levels in each stage of encoder network. Pyramid edge extraction module (PEE) is proposed for aggregating boundary information with multiple granularities of current level firstly. Then a mini multi-task learning module (mini-MTL) is adopted for getting richer knowledge leveraging potential correlations and complementary features among related boundary detection and segmentation tasks. At last, the learned feature maps of mini-MTL further module realize the complementarity among different levels and refine high-level features and low-level features through cross feature fusion

module (CFF). Finally, in decoder part, we obtain the decoding features $D_i, i \in \{1, 2, 3, 4\}$ by aggregating the output feature maps from the ASPP module and encoding features of each stage in turn for final segmentation prediction.

B. Pyramid Edge Extraction Module

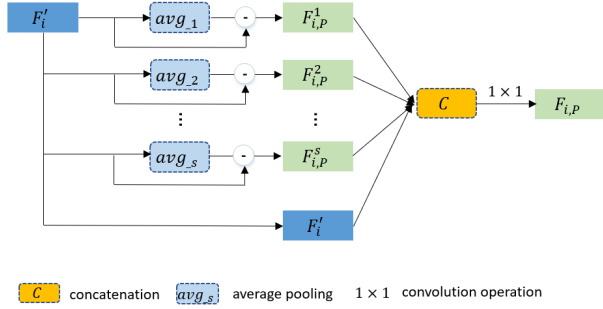


Fig. 3. The structure of the proposed pyramid edge extraction (PEE) module.

Edge of lesion region offers important information on the location of target objects. However, the boundary of the lesion area is usually complex and diverse. In order to obtain a robust boundary information supplement, we design a simple and effective pyramid feature extraction scheme for mining multi-granularity edge features in each stage of encoder network. As illustrated in Fig. 3, at first, we use the 1×1 convolution to squeeze the feature maps $F_i, i \in \{1, 2, 3, 4\}$ from last residual blocks in each stage of backbone and then employ them as the input for PEE module. It can be defined as follows:

$$F'_i = \mathcal{F}(F_i; \theta_i), \quad i \in \{1, 2, 3, 4\} \quad (1)$$

where F'_i denotes the reduced feature maps of each residual blocks, \mathcal{F} is the function of 1×1 convolution, and θ_i indicates the respective parameter. We obtain multiple granularities edge features by subtracting the value of average pooling with different sizes from its local convolutional feature maps. Without losing generality, we define S pooling operations:

$$F_{i,\mathcal{P}}^{(s)} = F'_i - avg_{-s}(F'_i), \quad s \in \{1, \dots, S\} \quad (2)$$

where $F_{i,\mathcal{P}}^{(s)}$ denotes the edge features of current i th stage with the s th pooling operation, and avg_{-s} is the corresponding average pooling operation. In order to integrate the obtained pyramid edge features, we aggregate them with the features of current stage by concatenating them together, and merge them using a 1×1 convolution operation.

$$F_{i,\mathcal{P}} = \mathcal{F}(C(F_{i,\mathcal{P}}^{(1)}, \dots, F_{i,\mathcal{P}}^{(S)}, F'_i); \theta_{i,\mathcal{P}}) \quad (3)$$

where \mathcal{C} refers to the concatenation process. $F_{i,\mathcal{P}}$ is the output feature maps of PEE module at current stage of encoder network. $\theta_{i,\mathcal{P}}$ represents the corresponding parameter. Such a multiple granularities edge features extraction design offers a powerful way to efficiently enhance the representation ability of the corresponding level. By extracting and

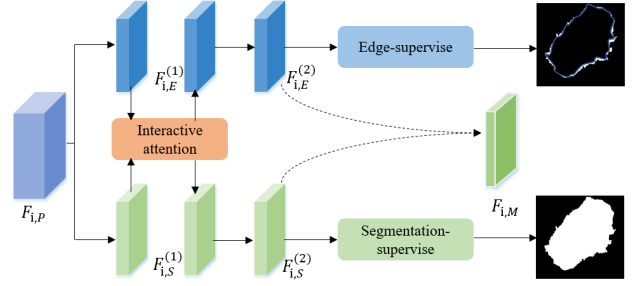


Fig. 4. The design of proposed mini multi-task learning (mini-MTL) module, which consists of two task-specific branches and an interactive attention layer.

integrating boundary information of different granularities, the edge features are effectively improved and noise is suppressed. Subsequently, the output maps are fed to the mini multi-task module to promote extraction of finer features.

C. Mini Multi-Task Learning Module

Naturally, additional knowledge from object edge can help to precisely identify the shape of the target, and the semantic segmentation and edge detection have strong complementary. Based on this idea, we propose a mini multi-task learning module (mini-MTL) embedded in each stage for jointly learning to segment object masks and detect lesion boundaries without introducing much parameters. The mini-MTL module aims at yielding performance gains by leveraging potential correlations among these related tasks. Fig. 4 displays the architecture of our proposed mini-MTL network. Each multi-task networks contains two components: the task-specific branch and an interactive attention layer. Specifically, each branch has two convolutional layers and an upsampling layer. The convolutional layer is used for encoding task-related features and upsampling layer is employed for obtaining the corresponding prediction mask. In i th stage, the feature maps $F_{i,\mathcal{P}}$ of PEE module are as the input of two subnetworks for further extracting task-related features simultaneously.

$$\begin{aligned} F_{i,\mathcal{E}}^{(l)} &= \mathcal{F}(F_{i,\mathcal{E}}^{(l-1)}; \theta_{i,\mathcal{E}}^{(l)}) \\ F_{i,\mathcal{S}}^{(l)} &= \mathcal{F}(F_{i,\mathcal{S}}^{(l-1)}; \theta_{i,\mathcal{S}}^{(l)}) \end{aligned} \quad (4)$$

where $F_{i,\mathcal{E}}^{(l)}$ and $F_{i,\mathcal{S}}^{(l)}$ indicate the feature maps extracted from l th convolutional layer of edge sub-network and the segmentation sub-network, $l \in \{1, 2\}$. In particular, $F_{i,\mathcal{E}}^{(0)}$ and $F_{i,\mathcal{S}}^{(0)}$ denote the feature $F_{i,\mathcal{P}}$ of PEE module. \mathcal{F} is the function of 3×3 convolution with respective parameters $\theta_{i,\mathcal{E}}^{(l)}$ and $\theta_{i,\mathcal{S}}^{(l)}$ respectively. The interactive attention (IA) is designed in the first convolutional layer for mining the interactive information from different tasks. As illustrated in Fig. 5, in order to integrate the effective information from other task, we design one simple yet effective interactive attention integration method assigning different weights to various regions. Specifically, take the edge features integration as an example, we first use the sigmoid function to generate a weight mask which indicates the important positions of current edge feature maps $F_{i,\mathcal{E}}^{(1)}$. Then the reverse attention weight is generated by subtracting

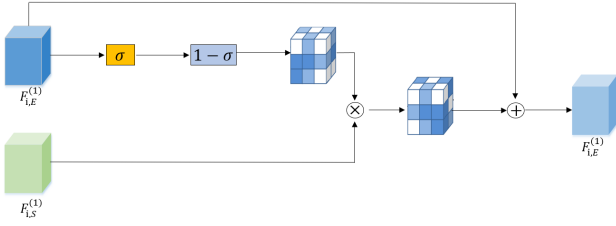


Fig. 5. The illustration of proposed interactive attention method. The figure shows the example of edge features integration by IA.

the weight mask from one. At last, we can selectively send the useful information from the segmentation feature to current edge feature $F_{i,S}^{(1)}$ by an element-wise product operation. Similarly, the feature integration on segmentation branch is similar to this. Thus, the whole interactive process can be formulated as follows:

$$\begin{aligned} F_{i,E}^{(1)} &= F_{i,E}^{(1)} + (1 - \sigma(F_{i,E}^{(1)})) \otimes F_{i,S}^{(1)} \\ F_{i,S}^{(1)} &= F_{i,S}^{(1)} + (1 - \sigma(F_{i,S}^{(1)})) \otimes F_{i,E}^{(1)} \end{aligned} \quad (5)$$

where σ denotes the sigmoid activation function to produce a filter mask. \otimes represents the element-wise product. The proposed IA module is based on gated mechanism without addition parameters, and in this manner the information from different tasks is delivered effectively by IA module. Moreover, the useful information can be regulated to the right place through attention and useless message can also be suppressed on both the sender and receiver sides simultaneously. After two convolution layers and interaction layer, we get more abundant context representation $F_{i,M}$ of current stage through aggregating the features from these two tasks:

$$F_{i,M} = \mathcal{F}(\mathcal{C}(F_{i,E}^{(2)}, F_{i,S}^{(2)}); \theta_{i,M}) \quad (6)$$

where \mathcal{F} and \mathcal{C} refer to the 1×1 convolution and concatenation operation respectively. $\theta_{i,M}$ denotes the parameter. For effectively multi-task learning, we jointly supervise and learn two branches together with IA module in an end-to-end manner. Here, both the edge map and segmentation map are all the binary representation of the outlines of objects and object classes. we expect the output of the edge and segmentation prediction of mini-MTL module to approximate the ground-truth masks respectively (represented as G_E and G_S) by minimizing the loss:

$$\begin{aligned} L_{i,E} &= BCE(P_{i,E}, G_E) \\ L_{i,S} &= BCE(P_{i,S}, G_S) \end{aligned} \quad (7)$$

where $BCE(\cdot, \cdot)$ means the binary cross-entropy loss function with the following formulation:

$$BCE(P, G) = - \sum_j^N (G_j \log P_j + (1 - G_j) \log(1 - P_j)) \quad (8)$$

where P_j and G_j indicate the j th pixel of predicted boundary (segmentation) maps $P_{i,E}$ ($P_{i,S}$) and ground-truth masks of

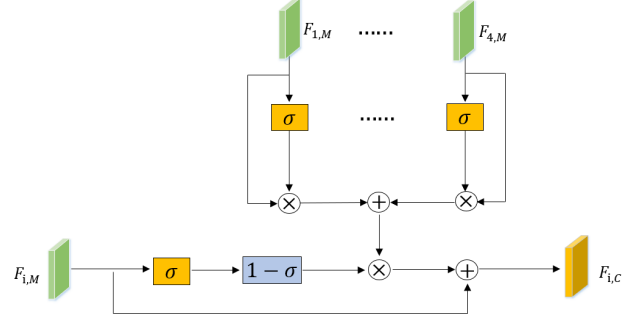


Fig. 6. The design of proposed cross feature fusion (CFF) module.

object boundary (segmentation) G_E (G_S), respectively. N represents the number of pixels. Thus the total loss in the i th mini-MTL module could be denoted as:

$$L_{i,M} = L_{i,E} + L_{i,S} \quad (9)$$

This joint learning helps to preserve the fine details near object boundaries. With this module, BA-Net is able to generate more accurate and better boundary-adherent features at current stage.

D. Cross Features Fusion Module

In encoder network, the low-level features are rich in spatial details and the high-level features have more abundant semantic information [42]. In order to utilize the complementary both spatial structural details and semantic information, we propose cross feature fusion module (CFF) to selectively aggregate features with different levels and refine both high-level and low-level feature maps. As shown in Fig. 6, for i th feature maps $F_{i,M}$, CFF adaptively selects complementary components from multiple input features $F_{j,M}, j \neq i$ by following attention mechanism:

$$F_{i,C} = F_{i,M} + (1 - \sigma(F_{i,M})) \otimes \left[\sum_{j \neq i} \sigma(F_{j,M}) \otimes F_{j,M} \right] \quad (10)$$

where σ denotes the sigmoid activation function. \otimes denotes the element-wise product. Therefore, information from different levels is integrated effectively by CFF module and can effectively avoid introducing too much redundant information. Therefore, we can get rich contextual features $F_{i,C}$ preserving rich details as well as semantic information at each stage through three modules in series which is used for decoder process.

E. Decoding and Optimization

By cascaded three modules at each stage of encoder network, richer context and fine-grain features of each stage are encoded. In decoder network, we obtain the decoding features $D_i, i \in \{1, 2, 3, 4\}$ by aggregating the output feature maps F_A from the ASPP module and encoding features of each stage in turn for final segmentation prediction:

$$D_i = \begin{cases} \mathcal{F}(\mathcal{C}(D_{i+1}, F_{i,c}); \theta_{i,D}), & i \in \{1, 2, 3\} \\ \mathcal{F}(\mathcal{C}(F_A, F_{i,c}); \theta_{i,D}), & i = 4 \end{cases} \quad (11)$$

where D_i denotes the decoding features at each stage, \mathcal{F} is the function of 1×1 convolution, and $\theta_{i,D}$ indicates the respective parameter of decoder part.

The supervision of the whole network adopts the standard binary cross-entropy loss for minimize the error between output from decoder network and ground truth. During the end-to-end training process, the total loss function is defined with joint losses of multi-task module as follows:

$$\min L_D + \sum_i \lambda_i L_{i,\mathcal{M}} \quad (12)$$

where L_D denotes the decoder loss and λ_i indicates the balance parameters. The λ_i is empirically set as 1.0. The boundary information participates in updating and guiding the generation of final segmentation prediction through loss, which makes the whole network aware of object boundary and refines the result.

IV. EXPERIMENTS

A. Materials

To evaluate the effectiveness of our method, we conduct experiments on five medical image datasets with various modalities, including dermoscopy images, endoscopic procedures images, X-ray images and retinal fundus images. Details of these datasets are briefly introduced as follows: **ISIC-2017** [43] includes a training set with 2000 annotated dermoscopy images and a total number of 600 images for testing. The image size ranges from 540×722 to 4499×6748 . **Kvasir-SEG** [44] consists of 1000 polyp images and their corresponding ground truth polyp masks annotated by expert endoscopists. **CVC-ColonDB** [52] contains 380 colonoscopy images coming from 15 short colonoscopy video sequences with size 574×500 . **SZ-CXR** [45], [46] is collected by Shenzhen No.3 Hospital in Shenzhen, China. The dataset contains 566 X-rays images with respective annotations. **RIM-ONE-R1** [53] is composed of 169 full retinal fundus images which has been annotated by five different experts.

B. Reference Model

In our work, we compare our proposed BA-Net with six previous state-of-the-art image segmentation methods, including FCN [25], U-net [24], MultiResUNet [47], AG-net [48], CE-Net [23] and Deeplabv3 [34]. For fair comparisons, the segmentation maps of these methods are generated by the original code released by the authors or directly provided by them. Moreover, all experiments use the same data preprocessing and the predicted segmentation masks are evaluated with the same evaluation metrics.

C. Evaluation Metric

To quantitatively evaluate the performance of the proposed BA-Net, in this paper, five widely accepted metrics for medical image segmentation task are computed as evaluation criteria. They include Dice Similarity Coefficient (DI), Jaccard Index (JA), Accuracy (AC), Sensitivity (SE) and Specificity (SP). The details are defined as follows:

$$\begin{aligned} DI &= \frac{2 \cdot TP}{2 \cdot TP + FN + FP}, & SE &= \frac{TP}{TP + FN} \\ JA &= \frac{TP}{TP + FN + FP}, & SP &= \frac{TN}{TN + FP} \\ AC &= \frac{TP + TN}{TP + FP + TN + FN} \end{aligned} \quad (13)$$

where TP , TN , FP and FN represent the number of true positives, true negatives, false positives and false negatives, respectively. They are all defined on the pixel level. Among these metrics, JA mainly reflects the overlapping pixels between estimated and ground-truth masks, which is the most important evaluation metric for segmentation task. In our work, we mainly rank the performance by JA.

D. Implementation Details

Training setting: The proposed framework is implemented based on the Pytorch 1.0 framework and is trained with one NVIDIA TITAN X GPUs. In the model, we use the standard stochastic gradient descent optimizer to train the whole end-to-end network with 0.9 momentum. The backbone of encoder network is based on ResNet-101 pre-trained on ImageNet. ReLU is chose as the default activation function. We set the initial learning rate as 10^{-4} and employ the ‘‘poly’’ learning rate policy for all experiments similar to [34]. The learning rate is multiplied by $(1 - \frac{iter}{total_iter})^{0.9}$ after each iteration, eventually terminated at 200 epochs. All the training data is divided into mini-batches for network training, the mini-batch size is set as 8 during the training stage. In each stage of encoder network, all feature maps are firstly reduced to 128 channels with convolution operation before PEE module. In our PEE module, we use the 5×5 and 7×7 average pooling operation for the first two blocks, and 3×3 and 5×5 pooling kernel are employed in last two blocks. All these edge feature maps remain the same size through padding operation. Finally, the output of the decoder is bilinearly interpolated to the same size as the input image directly.

Data preprocessing: In order to enhance diversity of training samples, several common data augmentation and sampling strategies are utilized before training the model. For the sake of fairness, the input preprocessing of all models remains constant. The input images are randomly flipped, rotation and center cropping to perform data augmentation. The scale used in cropping is from 50% to 100% and random rotation in the range of -10 to 10 degrees for original images. Both the horizontal and vertical flip operation with a preset probability are employed in our work. And due to the various size of the image in the datasets, all input images are uniformly resized into a resolution of 256×256 for training and testing.

TABLE I
SEGMENTATION PERFORMANCE ON FOUR BENCHMARK DATASETS

Methods	ISIC-2017					Kvasir-SEG					CVC-ColonDB					SZ-CXR				
	AC	DI	JA	SE	SP	AC	DI	JA	SE	SP	AC	DI	JA	SE	SP	AC	DI	JA	SE	SP
FCN [25]	93.9	84.1	75.2	82.2	97.0	96.7	85.9	78.9	87.5	98.1	98.5	91.4	74.0	80.0	99.6	96.7	92.9	86.9	92.0	98.2
U-net [24]	93.3	85.2	76.5	84.5	97.3	96.5	85.4	78.6	86.9	98.2	98.6	83.0	75.7	81.6	99.6	98.1	96.1	92.0	95.4	99.0
MultiResUNet [47]	93.6	85.2	76.8	83.9	96.8	96.8	87.0	80.5	88.5	98.2	98.5	82.8	75.6	86.9	99.1	98.1	96.0	92.4	94.6	99.3
AG-net [48]	93.5	85.3	76.9	83.5	97.4	97.2	88.1	81.8	88.8	98.1	98.9	84.9	76.1	83.7	99.5	98.1	96.1	92.5	95.6	99.0
Deeplabv3 [34]	94.0	86.4	78.3	85.9	96.9	97.3	89.2	83.5	90.4	98.2	99.2	92.0	86.4	92.0	99.5	98.0	95.8	92.2	95.3	98.9
CE-Net [23]	94.0	86.5	78.5	86.9	95.6	97.5	89.3	83.5	90.9	98.2	99.1	92.0	86.3	91.2	99.6	98.1	96.1	92.6	95.3	99.1
ours	94.7	88.2	81.0	89.9	96.4	97.7	91.1	86.1	90.8	98.6	99.3	93.7	88.4	93.6	99.6	98.2	96.2	92.8	95.8	99.0

TABLE II
SEGMENTATION PERFORMANCE ON ISIC-2017 DATASET

Methods	Averaged evaluation metrics (%)				
	AC	DI	JA	SE	SP
Team-Mt.Sinai (*1)	93.4	84.9	76.5	82.5	97.5
Team-NLP LOGIX (*2)	93.2	84.7	76.2	82.0	97.8
Team-BMIT (*3)	93.4	84.4	76.0	80.2	98.5
Team-BMIT (*4)	93.4	84.2	75.8	80.1	98.4
Team-RECOD Titans (*5)	93.1	83.9	75.4	81.7	97.0
DDN [36]	93.9	86.6	76.5	82.5	98.4
SLSDeep [50]	93.6	87.8	78.2	81.6	98.3
SegAN [37]	94.1	86.7	78.5	-	-
Chen et al. [51]	94.4	86.8	78.7	-	-
MB-DCNN [49]	-	87.8	80.4	-	-
ours	94.7	88.2	81.0	89.9	96.4

note: The *-number indicates the rank of that method in original ISIC-2017 challenge.

E. Comparisons with the State-of-the-Art

1) *Dermoscopy image dataset:* Skin lesion segmentation within dermoscopy images is very useful for automated melanoma diagnosis, especially for early protection and treatment. As shown in Table I, our BA-Net attains the highest DI of 88.2% and the best JA of 81.0% than other state-of-the-art methods. As for JA, our approach is noticeably improved from 78.5% to 81.0% on the test set compared to the best competitor CE-Net. In comparison with the results of classical FCN and U-net, our work exceeds them 5.8% and 4.5% on metric JA. Moreover, compared with the latest MultiResUNet and AG-net network, about 4.2% and 4.1% gains are obtained by our method, respectively. The above results suggest that the scheme design of cascaded three modules effectively extracts the richer context and fine-grain features for lesion recognition.

We also compare our BA-Net with top five methods submitted to the ISIC-2017 skin lesion segmentation challenge. As shown in Table II, our method achieves the best performance on the benchmark, outperforming the best result of the competition with 4.5% improvement on JA. In addition, we also report the other five recently published methods for skin lesion segmentation on the same test set. It is observed that our BA-Net has obvious advantages compared with the competitive published benchmarks. In particular, BA-Net outperforms the state-of-the-art MB-DCNN by 0.6% on JA. Compared with

MB-DCNN using mutual bootstrapping model with three stages, we have achieved outstanding performance in a simple and unified segmentation framework.

2) *Endoscopic image dataset:* Polyps are predecessors to colorectal cancer and therefore early treatment can help clinicians conduct risk screening and further diagnosis. Our method also achieves the best segmentation performance on both Kvasir-SEG and CVC-ColonDB datasets compared with the above state-of-the-art methods. Table I shows the comparison among them. For the Kvasir-SEG dataset, we divide the Kvasir-SEG including 800 images for training and 200 images for testing. Our BA-Net attains JA of 86.1%, which outnumbers the second competitor CE-Net and the Deeplabv3 by 2.6%. The images in CVC-ColonDB dataset from 15 different videos, where 304 images are used for training and 76 images for testing in our work. From the table, we can find that our method consistently outnumbers other state-of-the-art architectures and shows more performance gain on polyps segmentation. It implies that our method achieves the effective extraction of the same level features and integration of cross level features for improving the segmentation performance.

3) *X-ray image dataset:* We apply our BA-Net to segment lung organ in 2D X-ray image. In this dataset, we randomly separate the original 566 images into 452 training samples and 114 testing samples. The detailed experiment results are presented in Table I. The BA-Net achieves state-of-the-art performance in most metrics. The proposed method reaches an overall JA of 92.8%. In comparison with the results of classical FCN, the JA increases by 5.9%. The above results indicate our model with three modules substantially boosts the segmentation performance.

4) *Retinal fundus image Dataset:* We report the performance of our method for optic disc segmentation from retinal fundus images. The 169 images from RIM-ONE-R1 dataset are randomly separated to training and test sets with the ratio of 8 : 2. The RIM-ONE-R1 dataset contains five independent annotations from five experts. We compare the test results with the corresponding five expert marks, as shown in Table III. From the table, we can observe that our BA-Net achieves the best results both on single evaluation and overall average in JA. It further supports that our proposed PEE, mini-MTL and CFF modules are beneficial for medical image segmentation.

5) *Qualitative Evaluation:* Fig. 7 shows some examples of segmentation masks generated by our BA-Net as well as

TABLE III
SEGMENTATION PERFORMANCE ON RIM-ONE-R1 DATASET

Methods	Averaged evaluation metrics—JA (%)					Overall
	E1	E2	E3	E4	E5	
FCN [25]	92.1	90.6	91.1	90.1	90.9	90.9
U-net [24]	92.6	90.9	91.3	90.6	91.7	91.4
MultiResUNet [47]	93.1	91.6	92.5	91.5	92.9	92.3
AG-net [48]	92.9	91.4	92.1	90.3	92.1	91.8
Deeplabv3 [34]	93.2	91.7	92.4	91.9	93.2	92.5
CE-Net [23]	93.8	92.0	92.6	91.7	93.4	92.7
ours	93.9	92.4	92.6	92.1	93.7	93.0

note: The E-number indicates the different experts.

other state-of-the-art methods. We can see that objects can be highlighted well with accurate location and details by the proposed method. Meanwhile, similar background regions and noise are suppressed more thoroughly. From the row of 1 to 4 in Fig. 7, some methods only segment a part of the lesions and many other non-foreground information also get response. In comparison, our method performs very well. In the 4th and 5th rows, when dealing with lesion boundary of similar colors, other methods lack response to target lesions in various degrees, while our method can obtain the clear result. From the row of 6 to 9 in Fig. 7, we can observe that many methods usually extract coarse region mapping on the whole, but the details of the target area are missing due to noisy backgrounds or lower image contrast. In our method, target areas are more accurately located with the help of effective context representations. Overall, results show that the proposed BA-Net effectively mines the context of same level and uses the underlying correlations among features of distinct levels for capturing more abundant feature embeddings, which indicates the ability of processing the fine structures and rectifying errors.

F. Ablation Study

To validate the effectiveness of different components of the BA-Net, several ablation experiments are conducted on the ISIC-2017 and Kvasir-SEG benchmark datasets. The ablations results are shown in Table IV.

1) *Effectiveness of the BA-Net*: The proposed BA-Net designs three cascaded module in each stage of encoder network for refining and purifying the context, which achieves the outstanding segmentation performance. To justify the effectiveness, we compare the corresponding results by removing each module (denoted as “ours w/o PEE”, “ours w/o MTL” and “ours w/o CFF”) respectively.

As shown in Table IV, we can observe that the designed three modules greatly improve the segmentation performance compared with “baseline”. For these two datasets, the JA increases by 3.1% and 2.6% respectively, which indicates the usefulness of these modules for segmentation. After removing the PEE module from our model, the JA declines by 1.0% and 1.6% for both datasets due to the lack of the capability to utilize edge features with multiple granularities. Similarly, without using the mini-MTL module, the ability of the model

to make use of the complementarity of boundary information for segmentation is greatly weakened. This is also clearly reflected in the change of two data sets in JA, which decreases by 0.8% and 1.3%. Compared with the model only employing the PEE and mini-MTL modules, we can find that the application of CFF yields a 1.1% and 0.7% improvement of JA for these two tasks, respectively. It suggests that the CFF effectively bridges the features of different levels, and effectively integrates them for feature extraction at the current level. In fact, we find that any combination of two modules also brings large gains compared with “baseline”. When the network utilizes the PEE and mini-MTL module, the context of current level is fully utilized and mined. And the CFF module is employed for further integrating refined contextual information after PEE and mini-MTL modules from cross levels. From Table IV, by fully extracting the features of each level and reasonably integrating the information of different stages, our model has achieved excellent results.

2) *Effectiveness of the Interactive Attention*: In addition, we also investigate the effectiveness of the interactive attention in mini-MTL module. We conduct the experiment removing the IA of mini-MTL module and compare it with our original model. As shown in Table IV, for the no-interaction model, it gets the lower score of JA on two datasets and the JA declines by 0.7% and 1.1%. It implies that the IA between the edge sub-network and segmentation sub-network plays a great role in generating better representation for final segmentation prediction. As we expect, the IA fully considers the effect of different branches in mini-MTL module and effectively delivers the message to each other.

3) *Visualization comparison of ablations*: Fig. 8 shows the segmentation performances with different ablations. We find that ours can better highlight the objects without introducing background regions, and the visualization results of ours are better than any of w/o PEE, w/o MTL, w/o CFF and w/o IA. Specifically, the baseline model can extract coarse region contour on the whole. But for some images with low contrast and the image with unclear boundary, its segmentation accuracy is not unsatisfactory. From the comparison between module ablations and ours, with the help of PEE, our model obtains richer boundary information guidance, and the mini-MTL module makes full use of the complementarity of boundary information to segmentation. Meanwhile, the CFF module provides cross-level feature interaction, which enables the model to capture more overall context at each stage. In addition, the IA in mini-MTL enhances the delivery of effective information from two branches. This further demonstrates that the introduction of three modules designed in our work effectively enhances the segmentation performance.

V. CONCLUSION

In this paper, we present a boundary-aware context neural network, which produces richer contextual information for medical image segmentation. In our work, three cascaded modules are proposed. In each stage of encoder network, the pyramid edge feature extraction module extracts multiple granularities edge features, and then the mini multi-task learning

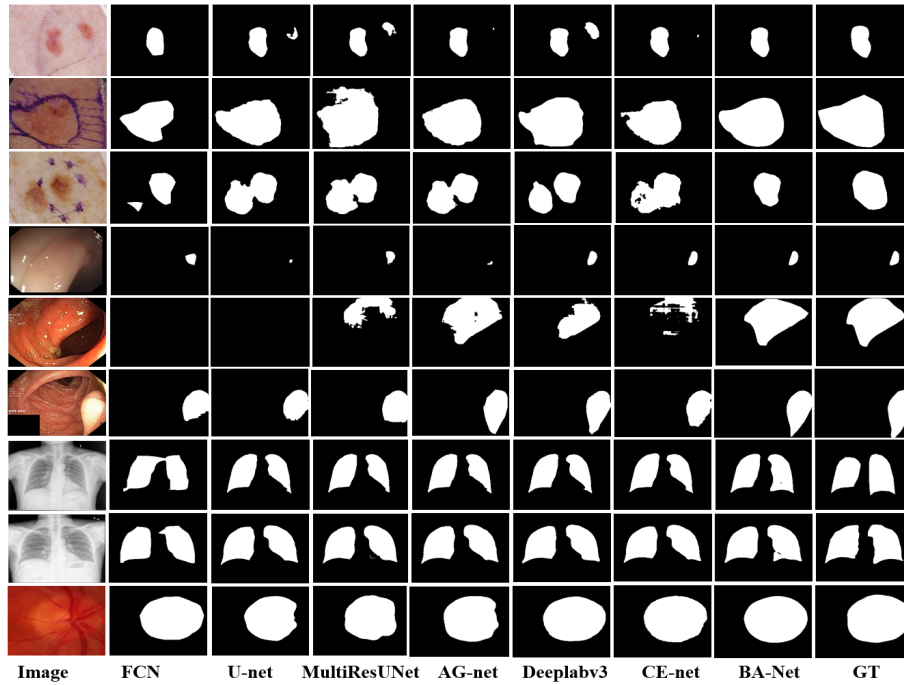


Fig. 7. Visual comparisons to six state-of-the-art methods on different medical image datasets.

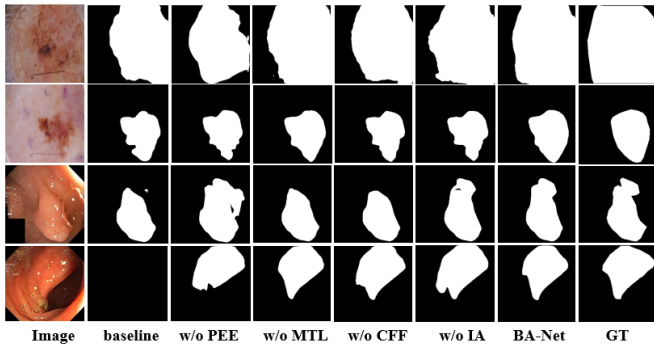


Fig. 8. Segmentation masks inferred with different ablations of our model on ISIC-2017 and Kvasir-SEG dataset. (w/o PEE: without PEE module, w/o MTL: without mini-MTL module, w/o CFF: without CFF module, w/o IA: without IA in mini-MTL module.)

TABLE IV
ABLATION ANALYSIS FOR THE PROPOSED NETWORK ON ISIC-2017 AND KVASIR-SEG DATASETS.

Methods	ISIC-2017		Kvasir-SEG	
	DI	JA	DI	JA
baseline	86.1	77.9	89.2	83.5
ours w/o PEE	87.6	80.0	90.1	84.5
ours w/o MTL	87.7	80.2	90.4	84.8
ours w/o CFF	87.3	79.9	90.8	85.4
ours w/o IA	87.8	80.3	90.6	85.0
ours	88.2	81.0	91.1	86.1

module effectively complements and enriches the context from the boundary detection branch through multi-task learning and designed interactive attention method. Finally, the network adaptively integrates the feature maps from different levels by the cross feature fusion module. By cascaded three mod-

ules, richer context and fine-grain features of each stage are obtained. Extensive comparative evaluations on five publicly available datasets are implemented, which has validated the superiority of the proposed method. Based on the outstanding performance of our work, we will extend our BA-Net to support 3D medical image segmentation tasks in the future.

REFERENCES

- [1] N. Sharma, and L. M. Aggarwal, "Automated medical image segmentation techniques," *J Med Phys.*, vol. 35, no. 1, pp. 3-14, Jan. 2010.
- [2] T. McInerney, and D. Terzopoulos, "Deformable models in medical image analysis: a survey," *Med. Image Anal.*, vol. 1, no. 2, pp. 91-108, Jun. 1996.
- [3] S. Hu, E. A. Hoffman, and J. M. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE Trans. Med. Imaging*, vol. 20, no. 6, pp. 490-498, Jun. 2001.
- [4] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, and H. M. Blau *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115-118, Feb. 2017.
- [5] A. Sanchez-Gonzalez, B. Garcia-Zapirain, D. Sierra-Sosa, and A. Elmaghraby, "Automatized colon polyp segmentation via contour region analysis," *Comput. Biol. Med.*, vol. 100, pp. 152-164, Sep. 2018.
- [6] M. Drozdal, G. Chartrand, E. Vorontsov, M. Shakeri, L. D. Jorjoc, and A. Tang *et al.*, "Learning normalized inputs for iterative estimation in medical image segmentation," *Med. Image Anal.*, vol. 44, pp. 1-13, Feb. 2018.
- [7] J. Jiang, P. Trundle, and J. Ren, "Medical image analysis with artificial neural networks," *Comput. Med. Imaging Graph.*, vol. 34, no. 8, pp. 617-631, Dec. 2010.
- [8] M. Silveira, J. C. Nascimento, J. S. Marques, A. R. S. Marçal, T. Mendonça, and S. Yamauchi *et al.*, "Comparison of segmentation methods for melanoma diagnosis in dermoscopy images," *IEEE J. Sel. Top. Signal Process.*, vol. 3, no. 1, pp. 35-45, Feb. 2009.
- [9] O. Acosta, A. Simon, F. Monge, F. Commandeur, C. Bassirou and G. Cazoulat *et al.*, "Evaluation of multi-atlas-based segmentation of CT scans in prostate cancer radiotherapy," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Jun. 2011, pp. 1966-1969.
- [10] M. Chen, E. Helm, N. Joshi, and M. Brady, "Random walk-based automated segmentation for the prognosis of malignant pleural mesothelioma," in *Proc. IEEE Int. Symp. Biomed. Imaging*, Jun. 2011, pp. 1978-1981.

- [11] X. Chen, B. M. Williams, S. R. Vallabhaneni, G. Czanner, R. Williams, and Y. Zheng, "Learning active contour models for medical image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 11632-11640.
- [12] A. V. Dalca, J. Guttag, and M. R. Sabuncu, "Anatomical priors in convolutional networks for unsupervised biomedical segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9290-9299.
- [13] L. Yu, H. Chen, Q. Dou, J. Qin, and P. A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imaging*, vol. 36, no. 4, pp. 994-1004, Dec. 2016.
- [14] V. Grau, A. U. J. Mewes, M. Alcaniz, R. Kikinis, and S. K. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Trans. Med. Imaging*, vol. 23, no. 4, pp. 447-458, Apr. 2004.
- [15] A. Tsai, W. Wells, C. Tempany, E. Grimson, and A. Willsky, "Mutual information in coupled multi-shape model for medical image segmentation," *Med. Image Anal.*, vol. 8, no. 4, pp. 429-445, Dec. 2004.
- [16] X. Chen, J. K. Udupa, U. Bagci, Y. Zhuge, and J. Yao, "Medical image segmentation by combining graph cuts and oriented active appearance models," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2035-2046, Apr. 2012.
- [17] A. S. Ashour, Y. Guo, E. Kucukkulahli, P. Erdogmus, and K. Polat, "A hybrid dermoscopy images segmentation approach based on neutrosophic clustering and histogram estimation," *Appl. Soft. Comput.*, vol. 69, pp. 426-434, Aug. 2018.
- [18] H. Ganster, P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imaging*, vol. 20, no. 3, pp. 233-239, Mar. 2001.
- [19] Y. He, and F. Xie, "Automatic skin lesion segmentation based on texture analysis and supervised learning," in *Proc. Asian Conf. Comput. Visi.*, Nov. 2012, pp. 330-341.
- [20] P. L. Bazin, and D. L. Pham, "Statistical and topological atlas based brain image segmentation," in *Proc. MICCAI*, Oct. 2007, pp. 94-101.
- [21] V. Cherukuri, V. K. BG, R. Bala, and V. Monga, "Deep Retinal Image Segmentation with Regularization Under Geometric Priors," *IEEE Trans. Image Process.*, vol. 29, pp. 2552-2567, Oct. 2019.
- [22] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 8543-8553.
- [23] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, and Y. Zhao *et al.*, "CE-Net: context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imaging*, vol. 38, no. 10, pp. 2281-2292, Oct. 2019.
- [24] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, Nov. 2015, pp. 234-241.
- [25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431-3440.
- [26] J. Carballido-Gamio, S. J. Belongie, and S. Majumdar, "Normalized cuts in 3-D for spinal MRI segmentation," *IEEE Trans. Med. Imaging*, vol. 23, no. 1, pp. 36-44, Jan. 2004.
- [27] D. H. Chung, and G. Sapiro, "Segmenting skin lesions with partial-differential-equations-based image processing algorithms," *IEEE Trans. Med. Imaging*, vol. 19, no. 7, pp. 763-767, Jul. 2000.
- [28] H. T. Nguyen, M. Worring, and R. V. Boomgaard, "Watersnakes: Energy-driven watershed segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 330-342, Mar. 2003.
- [29] J. Xie, Y. Jiang, and H. Tsui, "Segmentation of kidney from ultrasound images based on texture and shape priors," *IEEE Trans. Med. Imaging*, vol. 24, no. 1, pp. 45-57, Jan. 2005.
- [30] P. L. Bazin, D. and L. Pham, "Homeomorphic brain image segmentation with topological and statistical atlases," *Med. Image Anal.*, vol. 12, no. 5, pp. 616-625, Oct. 2018.
- [31] F. Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE Int. Conf. 3D Vis.*, Dec. 2016, pp. 565-571.
- [32] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, and D. Feng, "Dermoscopic image segmentation via multistage fully convolutional networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2065-2074, Jun. 2017.
- [33] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, and M. Aertens *et al.* "DeepIGeoS: a deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559-1572, Jul. 2018.
- [34] L. C. Chen, G. Papandreou, F. Schroff, H. Adam. (2017). "Rethinking atrous convolution for semantic image segmentation." [Online]. Available: <https://arxiv.org/abs/1706.05587>
- [35] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881-2890.
- [36] H. Li, X. He, F. Zhou, Z. Yu, D. Ni, and S. Chen *et al.*, "Dense deconvolutional network for skin lesion segmentation," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 527-537, Jul. 2018.
- [37] Y. Xue, T. Xu, and X. Huang, "Adversarial learning with multi-scale loss for skin lesion segmentation," in *Proc. IEEE Int. Symp. Biomed. Imaging*, May 2018, pp. 859-863.
- [38] Y. Zhou, X. He, L. Huang, L. Li, Z. Fan, and S. Cui *et al.*, "Collaborative learning of semi-supervised segmentation and classification for medical images," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2079-2088.
- [39] C. Chen, Q. Dou, H. Chen, J. Qin, and P. A. Heng, "Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, Jan. 2019, pp. 865-872.
- [40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770-778.
- [41] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, and S. Ma *et al.*, "Imagenet large scale visual recognition challenge", *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211-252, Apr. 2015.
- [42] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun, "Exfuse: Enhancing feature fusion for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 269-284.
- [43] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, and S. W. Dusza *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *Proc. IEEE Int. Symp. Biomed. Imaging*, May 2018, pp. 168-172.
- [44] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, and D. Johansen *et al.* "Kvasir-seg: A segmented polyp dataset," *Inter. Conf. Multimedia Modeling.*, Dec. 2020, pp. 451-462.
- [45] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, and Z. Xue *et al.*, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577-590, Feb. 2013.
- [46] S. Stirenko, Y. Kochura, O. Alienin, O. Rokovyi, S. Stirenko, and P. Gang *et al.*, "Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation," in *Proc. IEEE Inter. Conf. Electronics and Nanotechnology*, Oct. 2018, pp. 422-428.
- [47] N. Ibtihaz, and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74-87, Jan. 2020.
- [48] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, and B. Glocker *et al.* "Attention gated networks: Learning to leverage salient regions in medical images," *Med. Image Anal.*, vol. 53, pp. 197-207, Apr. 2019.
- [49] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification," *IEEE Trans. Med. Imaging*, Feb. 2020.
- [50] M. M. K. Sarker, H. A. Rashwan, F. Akram, S. F. Banu, A. Saleh, and V. K. Singh *et al.*, "SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks," in *Proc. MICCAI*, Sep. 2018, pp. 21-29.
- [51] 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 Int. Symp. Biomed. Imaging*, May 2018, pp. 1126-1129.
- [52] J. Bernal, J. Sánchez, and F. Vilarino, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166-3182, 2012.
- [53] F. Fumero, S. Alayón, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez, "Rim-one: An open retinal image database for optic nerve evaluation," in *Proc. IEEE CBMS*, 2011, pp. 1-6.