Biomedical Signal Processing and Control 71 (2022) 103165

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Region-to-boundary deep learning model with multi-scale feature fusion for medical image segmentation

Xiaowei Liu [a], Lei Yang [a,*], Jianguo Chen [a,*], Siyang Yu [b], Keqin Li [c]

[a] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[b] Department of Information Management, Hunan University of Finance and Economics, Changsha, China
[c] Department of Computer Science, State University of New York, New Paltz, NY 12561, USA

ARTICLE INFO

Keywords:
Boundary awareness
Region-to-boundary
Multi-scale features fusing
Medical image segmentation
Scale attention

ABSTRACT

Accurately locating and segmenting lesions, organs, and tissues from medical images are necessary prerequisites for disease diagnosis, monitoring, and treatment planning. Semantic segmentation refers to the classification of each pixel/voxel in two-dimensional or three-dimensional space, which is beneficial to clinical parameter measurement and disease diagnosis. Due to the diversity of features such as size, shape, location, and intensity, segmenting lesions or organs from medical images has always been a challenging worldwide topic. Especially for low-contrast medical images, boundary recognition is particularly difficult. In this paper, we propose a novel region-to-boundary deep learning model to provide a feasible solution to alleviate this problem. First, we use a U-shaped network with two branches behind the last layer, one of which generates the target probability map, and the other obtains the corresponding signed distance map. Secondly, with the help of the signed distance map and obtained multi-scale features, we focus on the boundary of the target lesions or organs to be segmented. Finally, we fuse the region and boundary features and acquire the final results. We conduct extensive experiments on two public data sets and compare with seven the representative methods. The results show that the proposed model is superior to the comparative methods in most evaluation metrics, especially boundary tracking.

1. Introduction

In recent years, with the development of Artificial Intelligence (AI) technology, especially Deep Learning (DL), there has been an unprecedented development in the high-level understanding of natural images. Some applications in certain areas, such as license plate recognition and face recognition, have been popularized in our daily lives. Many scholars have introduced various DL algorithms into medical image processing and analysis, and have achieved promising results [13,19].

In the process of clinical medical image analysis, radiologists usually use X-rays, B-ultrasound (B-US), Magnetic Resonance Imaging (MRI), Computed Tomography (CT) and other medical images to visualize organs, tissues and lesions, which can help doctors quickly make a correct diagnosis. Generally speaking, when doctors diagnose diseases based on medical imaging, they often need to locate the lesions and perform fine segmentation to facilitate subsequent measurement and further analyze the related indicators. This process can guide significance in disease assessment and surgical intervention. Currently, the correct diagnosis of diseases through medical imaging requires the full participation of experienced radiologists. The localization and segmentation of organs, tissues, and abnormal parts is an unavoidable link in the process of intelligent medical imaging.

There are two popular methods applied to medical image segmentation. One is based on a generative model, which uses a Generative Adversarial Network (GAN) to construct a generative model by using only health data [18]. The model projects disease images to healthy ones and roughly locate possible lesions. The other is based on a discriminative model, which is the most mainstream and practical choice [8]. In the following, we will discuss some critical methods that rely on discriminative models, especially the widely used U-shaped architecture.

Since 2015, Ronneberger et al. created a typical U-Net architecture for two-dimensional (2D) medical image segmentation [24]. A large number of experiments have confirmed the powerful strength of the U-Net model. Since then, many variants of U-Net models have been developed and applied to the field of medical image segmentation [15,20,22]. For example, in [20], Andriy et al. described an encoder-decoder-based architecture, which is a variety of U-Net and used for

* Corresponding authors.
E-mail addresses: liuxiaowei@hnu.edu.cn (X. Liu), jt_yl@hnu.edu.cn (L. Yang), jianguochen@hnu.edu.cn (J. Chen).

https://doi.org/10.1016/j.bspc.2021.103165
Received 1 June 2021; Received in revised form 12 August 2021; Accepted 7 September 2021
Available online 17 September 2021
1746-8094/

brain tumors segmentation. An additional decoder was interpolated into the bottleneck of the model, and it can be used to regularize the shared encoder. Fortunately, the achievement ranked first in the MICCA. In the same year, Isensee et al. proposed a NNU-Net model [15]. They did not design a complex structure, but focused on data preprocessing, results post-processing, and automatic hyper-parameter adjustment. Surprisingly, this model won the all-around champion in the Medical Segmentation Decathlon.[1] Even so, based on a large amount of literature and experiments, we cannot ignore the importance of network structure. For instance, Oktay et al. suggested Attention U-Net [22], a novel attention gate (AG) model for medical image segmentation.

Following U-Net, Oktay et al. proposed a new Attention Gate (AG) model for medical image segmentation [22]. Compared with the original U-Net architecture, the performance of the AG model has been significantly improved. Alom et al. designed two related models based on U-Net, called R2U-Net and Attention R2U-Net [1]. Gao et al. improved U-Net by employing an additive channel–spatial attention (ACSA) module in skip connection and applying multi-scale deep supervision to different layers of the decoding module, which made good progress in model performance [7]. Reza and Azad et al. respectively built the BCDU-Net [3] and MCGU-Net [2] models, in which two-way LSTM, densely connected convolution and Squeeze-and-Excitation (SE) block are correctly integrated into U-Net. The models also integrated with the corresponding U-Net and other latest technologies, and achieved good improvements. In [23], Qin et al. provided a nested U-shaped structure for salient target detection, in which the basic convolution block is replaced by a U-shaped sub-block. Unlike ordinary U-Net model, the U-shaped sub-block does not change the characteristic channel in up-sampling and down-sampling stages.

In practical applications, it is a good concept to consider boundary information in medical image segmentation. In [26], Wu et al. proposed a two-level neural network, where a center-sensitive mechanism was embedded into the global heat map to ensure the center of the tooth is accurately found. Then, in the local stage, a dense ASPP-UNet module was used to fine segment each single tooth. In [21], Andriy et al. introduced a fully CNN model with end-to-end boundary perception. By designing a special boundary branch supervised by the loss of edge perception, the kidney and kidney tumors can be reliably segmented from the 3D CT scan of the artery and abdomen. In [12], Hu et al. suggested a boundary-aware network for kidney and kidney tumor segmentation. The model uses a skip connection from the boundary decoder to the segmentation decoder to guide the segmentation process, especially for error-prone regions. In [29], Zhou et al. recommended a volume progressive lesion segmentation model, which uses a scale-invariant and boundary-aware deep convolution network to automatically segment 3D lesion volume from the 2D contour. Two additional studies designed dedicated edge-aware branches to capture richer boundary-aware context, and achieved good results by increasing their perception of boundary information [10,29].

In this paper, we focus on the boundary recognition of low-contrast medical images, and propose a novel region-to-boundary deep learning model to provide a feasible solution for medical image segmentation. The proposed model can not only ensure the overall segmentation performance, but also consider the boundary tracing. The main contributions of this work are summarized as follows:

- We propose a region-to-boundary deep learning model and divide the segmentation task into two stages. In the first stage, we classify all input pixels indiscriminately. In the second stage, we focus on the boundaries of all target medical tissues. Finally, we fuse the two intermediate results to get the final outcome.
- To locate the boundary features required in the second stage, we obtain the signed distance map of the segmented object at the same time, and get the boundary attention through a simple transformation $(1-|SDM|)$. With the boundary attention matrix, the subsequent refinement network will focus on edge pixels.
- We conduct extensive experiments on commonly used data sets and compare the proposed model with the state-of-the-art methods. The experimental results show that our model is superior to most comparative methods in terms of numerical evaluation criteria and vision, further verifying the effectiveness and feasibility of the proposed model.

The rest of the paper is organized as follows. Section 2 reviews the related work from the perspectives of U-Net, signed distance map, and visual attention mechanism. Section 3 describes the structure of our proposed region-to-boundary segmentation model and the related loss functions. Experimental results and performance evaluations are discussed in Section 4. Section 5 concludes the paper with a discussion of future work.

2. Related work

2.1. U-Net

U-Net is a typical U-shaped structure, including an encoder and a decoder, and has received extensive attention in recent years. By reducing the size of the feature map and increasing the channels, the encoder expects to extract high-level semantic features. Then, the decoder up-samples the output of the previous layer and reduces the number of channels layer by layer, which is symmetrical with the encoder. However, quite apart from this, there are also skip connections between the symmetrical layers of the encoder and decoder. 3D U-Net [5] and 2D U-Net [24] can be regarded as different versions of U-Net. 2D U-Net is not only suitable for 2D image segmentation, both have good segmentation performance for 3D medical images. In recent years, each of them has its wins or losses in different tasks in MICCAI.[2]

Since the publicity of the U-Net framework, a large number of improved versions have emerged, and excellent excellent have been achieved, especially in the application of medical image segmentation. For instance, in [14], the TernausNet model replaced the encoder of U-Net with VGG11, which is pre-trained on ImageNet. Fortunately, it stood out from 735 teams and won the first prize in the annual Carvana Image Masking Challenge.

In many cases, the improvement work can achieve good results through small changes based on the U-Net model. For example, the Res-UNet model in Ref. [27] integrated the residual connections based on U-Net. The residual connection refers to the skip link of the front and back layers. The Dense-UNet model in Ref. [9] adopted the dense connections The dense connection here means that the output of a certain layer in the sub-module is regarded as a part of the input of the subsequent layers. The input of a certain layer comes from the combination of part or all of the outputs of the previous layers.

In [13], Ibtehaz et al. proposed a MultiResUNet model with multi-residual modules, where the three outputs of three consecutive convolutions and the $3\times$ 3filter are spliced together as a combined feature map. Then, a residual connection is added to the model by using a 1×1 convolution. In [1], Alom et al. proposed a R2U-Net model. They combined the residual connection and the recurrent convolution to replace the corresponding sub-components on the U-Net model. The performance of the R2U-Net model was evaluated by several common data sets such as skin disease images, retina images, and lung images. Note that U-Net [22] introduces the attention mechanism into the standard U-Net. An attention module is inserted on the skip connection to re-adjust the outputs of the encoder, thereby generating gated signals to control the importance of different spatial features. This so-called importance

[1] http://medicaldecathlon.com/.

[2] http://www.miccai.org/.

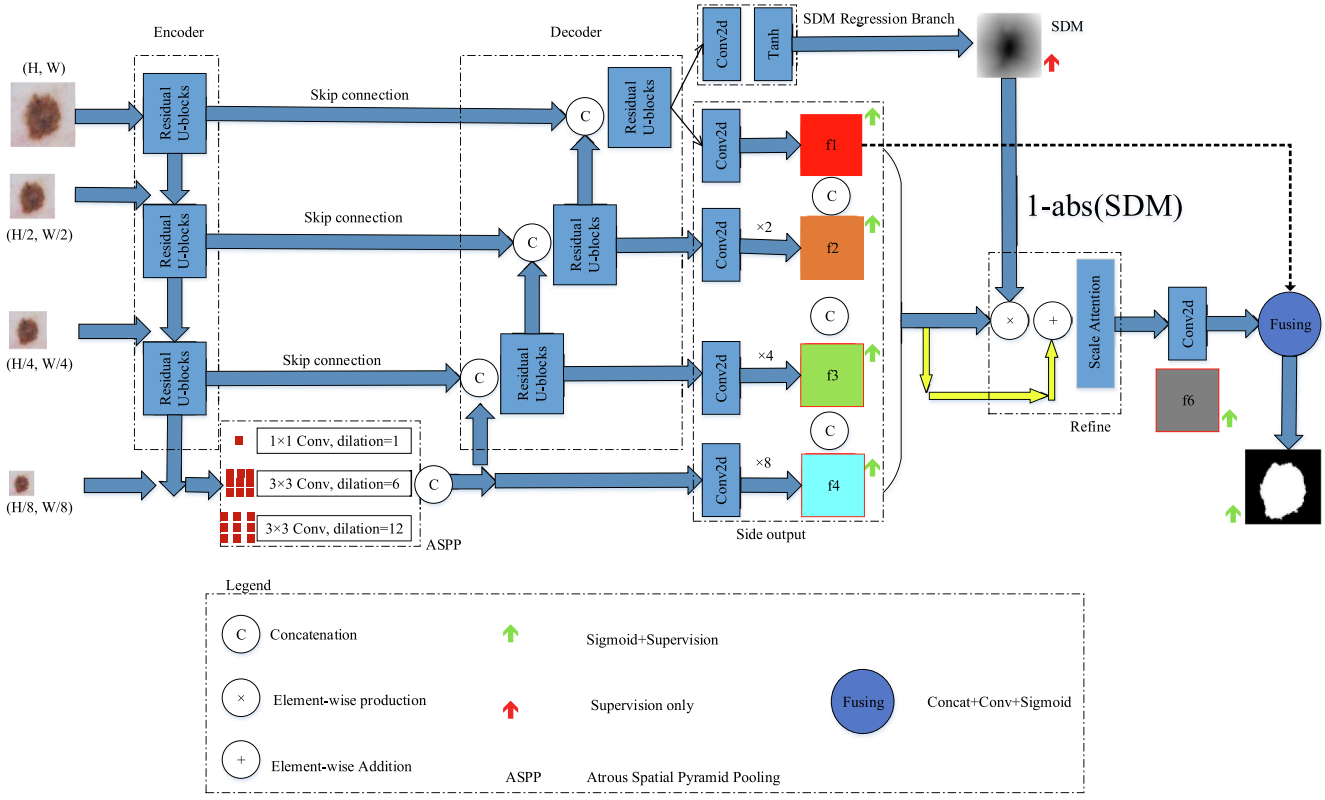


Fig. 1. Overall structure of the proposed region-to-boundary segmentation model. The proposed model consists of two parts, one is used for region segmentation, and the other is responsible for exploring the classification of pixels near the edge with the help of SDMs and multi-scale feature maps. The former is mainly composed of an encoder, an ASPP module, and a decoder. The encoder receives multi-scale inputs here. In the bottleneck, ASPP uses multiple receptive fields to refine image features. In the decoder, each layer has deep supervisions, and there is also an SDM regression branch to assist in obtaining boundary information. The latter mainly consists of a scale attention module and a fusing module. The scale attention module fuses the output features of different scales and the fusion module reconciles f1 and the output of scale attention modules.

means that the weights near the target are often larger than other regions, which is very useful for improving the accuracy of target segmentation. In [2], Asadi et al. proposed the BCDU-Net model, which is a hybrid of U-Net, bi-directional Long Short Term Memory (LSTM), dense convolution, and Squeeze Excitation (SE) block, and has gained amazing results.

2.2. Signed distance map

For the binary object segmentation mask, the corresponding distance map [4] can be acquired by calculating the distance from each pixel (marked as 1) in the target area to the nearest boundary pixel. This representation provides rich and powerful knowledge about the boundaries, shapes, and positions of segmented objects. In the same way, perform the above calculation on the background to obtain another distance map. The two distance maps take different signs, and then they are added by element. The result is a Signed Distance Map (SDM). Mathematically speaking, for a binary segmentation mask, SDM is usually defined in Eq. (1):

$$\phi(x) = \begin{cases} 0, & x \in \partial\Omega; \\ -\inf\limits_{y\in\partial\Omega} \|x-y\|_2, & x \in \Omega, \Omega \neq \varnothing; \\ +\inf\limits_{y\in\partial\Omega} \|x-y\|_2, & x \notin \Omega, \Omega \neq \varnothing; \\ 1, & \Omega = \varnothing, \end{cases} \tag{1}$$

where $\Omega = \{x_i | l_i = 1, i \in \mathscr{S}\}$ is the pixel set of foreground, x_i denotes any pixel, l_i represents the corresponding label, the index i is traversed the entire input image or the corresponding segmentation mask, and $\mathscr{S}$ is the index set. At the same time, we mark the boundary pixel set as $\partial\Omega$. The SDM defined above has no upper limit. Because it changes with the size of the segmented objects and the entire images, so it is often normalized first in use. Here, the middle two terms in this piece-wise function are normalized to the range $[-1, 1]$ by the Mini-Max method.

Generally, the value of an element in SDM implies the vector distance from any point to the nearest pixel on the boundaries. The negative sign represents the target region, and the positive sign represents the background region. SDM is also a general definition of level set functions, so we name it $phi(x)$. It should be noted that when there is no foreground, the value of each element in SDM will be set to 1.

2.3. Visual attention mechanism

Visual attention is a physiological mechanism that focuses on certain things and ignores most other information. If the background information is not filtered out by the attention mechanism, a person will be submerged by the large amount of visual information captured by the eyes. Inspired by this biological mechanism, computer vision directly benefits a lot. In most DL models, the attention mechanism is usually implemented as an add-on block, which can be easily plug and play to assign different weights to different regions. In addition to hard attention, soft attention is more widely used, which is also our concern.

Soft attention is like looking through a foggy glass window, and hard attention is like looking forward with a telescope. In the former case, we can see the whole image, but concentrate on certain places. In the latter case, we can only see a part of the world ahead, hopefully, the part most relevant to our task. Hard attention means that a pixel is either completely visible or completely invisible, which is equivalent to a 0–1 discrete problem. This non-differentiability directly leads to the

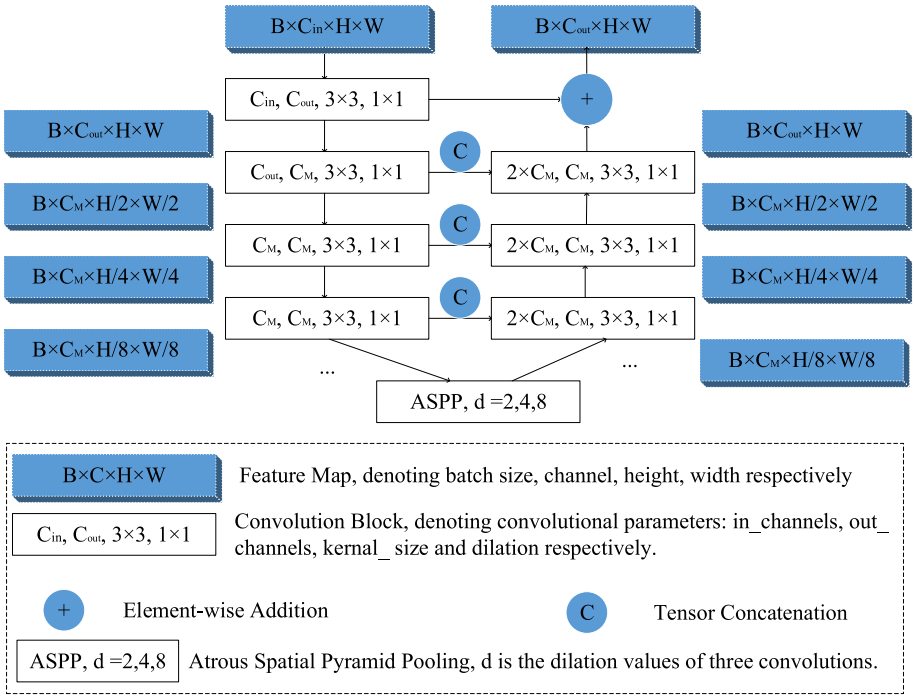


Fig. 2. Illustration of the Residual U-shaped block. Unlike the traditional U-Net, the channel of the feature maps will not change during up-sampling and down-sampling, but only the size will change. The output of the first convolutional block is connected to the end of the decoder through element-wise addition. This block can be expressed in parameter form as $RUB(C_{in}, C_M, C_{out})$, where C_{in}*and*C_{out} represent the input and output channels respectively, and C_M is the number of intermediate channels used in the inner layers.

difficulty of training. For the "clipping part" that only needs to be stored and manipulated, the calculation and memory requirements are very small. The soft attention level of each region, channel, time, or other dimension is expressed as a floating number from 0 to 1, which is a continuous problem. Due to its differentiability, soft attention is easy to train, but usually requires more memory and computation resources than hard attention.

Soft attention can act on space, channel, and time dimensions, respectively, while temporal attention mainly acts on sequence data such as videos and natural language. So for static vision, we usually only use spatial attention, channel attention, and their mixed mode. As for spatial attention, Spatial Transformer Networks (STNs) [16] are a classic case, which use spatial transformers to transform spatial information into another space and retain key features. The spatial transformer is essentially the realization of the spatial attention mechanism, because the trained spatial transformer can find out the area in the image that needs to be focused.

For channel attention, we can understand the principle of channel attention from the perspective of signal conversion. In signal analysis, any signal can be decomposed into a linear combination of sine waves. For example, in the Convolutional Neural Networks (CNNs), many new signals are generated in each channel through multiple convolutional kernels. That is, each input features is decomposed into multiple channels. In [11], Hu et al. focused on the relationship between the channels, and proposed a new type of structure called Squeeze and Excitation (SE). It is essentially a process of amplification and contraction. By multiplying the features of different channels with different weights, the attention of key channel regions can be improved.

For mixed attention, on the one hand, spatial attention ignores the information in the channel and treats the features in each channel equally, which limits its ability to extract features. On the other hand, the channel attention focuses on the global information of multiple channels, while ignoring the local information of each channel. To integrate the advantages and avoid the disadvantages of the two, mixed concerns have emerged. Dual Attention [6] and CBAM [25] are two typical examples of this attention mechanism. In dual attention, spatial attention and channel attention are executed in parallel, while in CBAM, the two attentions are connected in series.

3. Proposed method

In this section, we will manifest the structure of our proposed region-to-boundary segmentation model and the related loss functions.

3.1. Overall structure

The structure of the proposed region-to-boundary segmentation model is shown in Fig. 1. The u-shaped network on the left is used to classify all pixels equally. Among them, multiple instances of a sample are input into each layer of the encoder. In the decoder, feature maps of different scales are output from each layer. At the end of the decoder, there are two branches, one is to generate a segmentation map, and the other is generate a regression SDM. All the features from the decoder are first concatenated along channels, and then thrown to the next boundary-aware module, where we focus on the edges of segmented objects. These modules involved in the proposed model are detailed below.

3.2. Region segmentation

The task of region segmentation is completed by a U-shaped network, including an encoder and a decoder, as shown in Fig. 1. In the encoder, four instances of a sample with different scales are input into the network to obtain multi-scale features. As shown in Fig. 2, The convolutional block of each layer is composed of the Residual U-shaped Blocks (RUB), which is similar to the RSU structure in U2-Net [23].

The structure of $RUB(C_{in}, C_M, C_{out})$ is a typical U-shaped structure, as shown in Fig. 2. Among them, C_{in} and C_{out} represent the input and output channels respectively, and C_M is the number of channels in the RUB internal layer. It can be seen that the u-shaped block here is different from ordinary U-Net. Except for the first two basic convolutional blocks, the parameters *in_channel* and *out_channel* of other convolution blocks are both C_M. The basic convolutional block consists of a convolution, a batch normalization, and a ReLU function. The number of channels is not doubled during downsampling in the encoder, and the number of channels is not halved during upsampling in the decoder. That is, the size of the feature map is this block will never change. Last but not least,

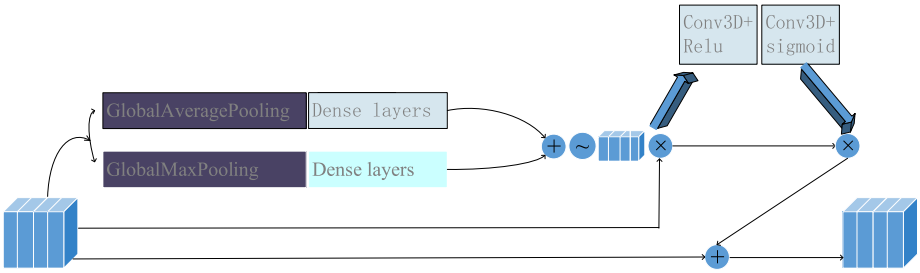


Fig. 3. Illustration of the Scale Attention with residual connections, where Channel Attention (CA) and spatial attention are chained together, and its input is a combination of resized feature maps with different scales obtained in the decoder.

there is a residual connection between the output of the first convolution block and the end of the decoder. At the bottleneck, we place the Atrous Spatial Pyramid Pooling (ASPP) block for feature extractions at different scales. Then, we balance the number of downsampling in the RUB layer according to the scale of the current layer.

In the decoding stage, the size and channels are gradually recovered upward. We use 1×1 convolution to change the number of channels to 1 on the right side of each layer. Then, we upsample the obtained single-channel features to the size of the first input. By performing the Sigmoid function, multi-scale deep supervision is added in the model. It is worth noting that all single-channel features that are not activated by the Sigmoid function are spliced together as the input of the next stage. At the end of the decoder, in addition to the segmentation branch, there is an auxiliary branch used to obtain SDM through regression. This operation will help subsequent modules to pay more attention to boundary/edge details.

3.3. Refined segmentation

Overall Process. After region segmentation, we collect all the feature maps of different sizes, including the four side outputs on the right side of the decoder, and then refine their boundaries. Specifically, these feature maps are first concatenated together along the channels. Secondly, the side features are enhanced by SDMs and sent to the scale attention blocks for boundary refinement. Finally, the two intermediate results are merged to obtain the result. We formalize this process as follows:

$$\mathbf{f}_{coarse} = Concat(f_1, f_2, f_3, f_4), \tag{2}$$

$$\mathbf{I}_{scale} = (2 - abs(SDM)) \times f_{coarse}, \tag{3}$$

$$\mathbf{f}_5 = SA(I_{scale}), \tag{4}$$

$$\mathbf{f}_{final} = conv(Concat(f_5, f_{coarse})), \tag{5}$$

$$\mathbf{O}_{final} = Sigmoid(f_{final}), \tag{6}$$

where f_1, f_2, f_3, f_4 are the output features on the right side of each layer in the decoder (as shown in Fig. 1). They have the same size and channel. For Eq. (2), we firstly use *Concat* to connect these features along the channel. Then, we use Eq. (3) to calculate the features with detail edge information, and input them to the Scale Attention (SA) block by Eq. (4). Finally, we simply fuse f_5 and f_{coarse} by Eqs. (5) and (6) and get the final output.

Scale Attention. Scale Attention (SA) is a special block used to fuse multi-scale features. As shown in Fig. 3, the first part is a channel-wise attention, which includes two parallel SE paths and generates a proportional coefficient in $[0, 1]$ for each channel. The second part is a spatial attention, which consists of two ordinary convolutions. The first reduces the number of channels of the feature maps, and the second further reduces the number of channels to 1. That is, these two attention modules create a spatial map by highlighting the regions of interest.

All in all, each of the two attention modules has a jump connection. The former is used for the feature fusion of different channels, and the latter is used for the spatial feature fusion of all channels. Last but not least, the module is to fuse multi-channel and spatial features by using enhancement rather than exclusion. Therefore, the fusion features and the input feature map are added at the end of the proposed model.

3.4. Loss functions

Similar to HED [28] and U2Net [23], we train our model in a deep supervision mode. The loss functions in region segmentation are different from the loss functions in fine segmentation. The former uses binary cross-entropy for segmentation, such as the L_2 loss for SDM regression. In contrast, the latter uses cross-entropy for segmentation, including dice loss and boundary loss. We define the loss functions used in our work.

Losses in Region Segmentation. As shown in Fig. 1, for region segmentation, four feature maps f_1, f_2, f_3, f_4 and their concatenation f_{coarse} (2) are used as deep supervision. Here, binary cross-entropy is hired to calculate the average cross-entropy over all pixels. Let Ω denote the domain of all pixels and y be the ground truth. Further, let $\hat{y}$be the predicted probabilities of each pixel, which comes from the output of the Sigmoid function, taking f_1, f_2, f_3, f_4, and f_{coarse} as input. Based on the above assumptions, our training loss for region segmentation is defined as:

$$L_k(\hat{y}, y) = \frac{1}{|\Omega|} \sum_{i \in \Omega} -y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \tag{7}$$

where $k = 0, 1, 2, 3, 4$ corresponds to the five supervised terms. In addition, we use L_2 loss between the predicted SDM and real SDM for the SDM regression. This term can be written easily as below Eq. (8):

$$L_{l2} = ||SDM_{pred} - SDM_{gt}||_2, \tag{8}$$

where SDM_{pred} denotes the predicted SDM and SDM_{gt} is the SDM calculated from labeled segmentation masks. So in general, the total loss in region segmentation is summarized as follows:

$$L_{region} = L_{l2} + \sum_{i=0}^{4} L_k, \tag{9}$$

where no weight is set up, means that the weight of all sub-losses is equal to 1.

Losses in Refined Segmentation. For refined segmentation, in addition to the Cross-Entropy (CE) loss (treating all pixels equally), we also use boundary loss (focusing on points around the boundary) to enhance the boundary perception. We define the former as:

$$L_{refine} = L_5 + L_6, \tag{10}$$

where L_5 means the CE loss between ground truth and the output of f_6

Table 1
Comparison results on the ISIC-2018 data set.

Method\ Metric	Dice	HD95	SP	SE	ACC	PC	JS
U-Net [24]	0.832 ± 0.179	4.702 ± 2.152	0.987 ± 0.040	0.784 ± 0.197	0.941 ± 0.090	0.934 ± 0.169	0.742 ± 0.206
Attention U-Net [22]	0.857 ± 0.156	4.804 ± 2.269	0.980 ± 0.049	0.856 ± 0.181	0.944 ± 0.090	0.900 ± 0.160	0.775 ± 0.192
R2U-Net [1]	0.867 ± 0.139	4.67 ± 2.147	0.971 ± 0.050	0.882 ± 0.150	0.949 ± 0.073	0.887 ± 0.159	0.786 ± 0.170
Attention R2U-Net [1]	0.857 ± 0.150	4.648 ± 2.010	0.986 ± 0.025	0.830 ± 0.183	0.946 ± 0.085	0.924 ± 0.138	0.771 ± 0.180
MSF-ACSA [7]	0.862 ± 0.152	4.459 ± 2.109	**0.987 ± 0.037**	0.825 ± 0.176	0.947 ± 0.086	**0.948 ± 0.134**	0.782 ± 0.186
BCDU-Net [2]	0.875 ± 0.142	4.625 ± 2.028	0.976 ± 0.039	0.884 ± 0.154	0.956 ± 0.059	0.900 ± 0.148	0.799 ± 0.169
U2-Net [23]	0.880 ± 0.145	4.371 ± 1.991	0.982 ± 0.048	0.889 ± 0.161	0.954 ± 0.081	0.913 ± 0.141	0.808 ± 0.171
Ours	**0.887 ± 0.133**	**4.360 ± 1.974**	0.979 ± 0.043	**0.892 ± 0.156**	**0.957 ± 0.069**	0.919 ± 0.129	**0.817 ± 0.168**

(as shown in Fig. 1), and L_6 is the CE loss between ground truth and O_{final} (as shown in Eq. (6).

Inspired by Kervadec et al. [17], we employ the boundary loss in our work:

$$\mathscr{L}_B(\theta) = \int_{\Omega} \phi_G(q) p_\theta(q) dq, \tag{11}$$

where p_θ is the predicted probability map of the input q, θ is model parameters, ϕ_G is a real SDM, which can be acquired readily from the labeled mask G. From Eq. (11), the boundary loss is a weighted sum of all elements of predicted probability maps, and ϕ_G acts as the weight tensor.

Total Losses. Ultimately, we define the final loss by combining the boundary loss with the standard regional loss:

$$\mathscr{L}_{final} = (1-\alpha) L_R(\theta) + \alpha \mathscr{L}_B(\theta), \tag{12}$$

where α is a coefficient weight that balances the boundary loss and the regional loss, as defined as:

$$\mathscr{L}_R = L_{refine} + L_{region} = \sum_{i=0}^{6} L_k + L_{l2}. \tag{13}$$

4. Experiments

4.1. Implementation details

We use PyTorch to implement the proposed DL framework, and conduct all experiments on a device equipped with XEON E5-2678V3 CPU, 32G RAM, and NVIDIA 2080Ti GPU with 11G memory. For hyper-parameter setting, the initial learning rate is set to $2e-4$, the weight decay is set to 0.0005, and the momentum is set to 0.99. In particular, for α in Eq. (12), we use a step-dependent Gaussian warming up function:

$$\beta(t) = \exp\left(-5\left(1 - \frac{t}{t_{max}}\right)^2\right), \tag{14}$$

to balance the regional loss and boundary loss, where t represents the current training step, and t_{max} means the maximum training step. In the final stage of our method, we fuse the results of boundary and region segmentation as the final output. The experimental results show that our framework is superior to the comparison methods in different evaluation metrics.

The experimental results show that the method is superior to the existing representative methods in the six evaluation indicators.

4.2. Data sets and evaluation metric

4.2.1. Data sets

We train and test our model on three widely used datasets: ISIC-2018,[3] a lung segmentation dataset,[4] and left atrial segmentation.[5]

ISIC 2018. The ISIC data set comes from 2018 ISIC challenges for three tasks: lesion segmentation, lesion attribute detection, and disease classification. We use the training data set of the first task, which contains 2594 images and the corresponding masks. To train, validate, and test our proposed framework, we divide it into three parts: 1815 images for training, 259 images for validation, and 520 images for testing. Due to hardware limitations and the size of each sample, we adjust the size of all images and corresponding masks to 256 × 256. There are no more data pre-processing operations, including spatial transformation and augmentation of color, brightness, and noise, except normalizing the samples before inputting them into the comparison models.

Lung Segmentation Data Set. The lung segmentation data set is a collection of CT images. It is introduced in the 2017 Kaggle Data Science Bowl's Lung Nodule Analysis (LUNA) competition to detect lung lesions in CT images. We first truncate the value of each 3D CT image to the range of [−512, 512], and then normalize them to [0, 1] through min–max normalization. We divide all slices containing lungs into a training set (70%) and a test set (30%), and adjust the size of each image to 512 × 512.

Left Atrial Segmentation Data Set. This data set includes 154 3D MRIs from patients with atrial fibrillation. The original MRI images are in grayscale, and the manual labels are presented by a binary mask. The size of the MRI images on the X-Y plane may vary from patient to patient, but the z-axis contains 88 fixed 88 slices. In our experiments, we slice 3D MRI images along the z-axis and re-sample them to 512 × 512 on the X-Y plane. To facilitate local testing, we only use 100 3D training data with labels. These 3D data are cut into 2D along the z-axis, and then divide them at a ratio of 7:3. The former is used for training and the latter is used for validation.

4.2.2. Evaluation metrics

For medical image segmentation, we use several related metrics for performance evaluation of the comparison methods, including Dice, HD95, Jaccard Similarity (JS), Accuracy (AC), Specificity (SP), Sensitivity (SE), and Precision (PC). When the validation loss remains unchanged for 12 consecutive epochs, we will stop model training. For the ISIC2018, LUNG data sets and LA segmentation, the maximum number of training epochs is set to 100.

4.3. Comparison experiments

In this section, we compare the proposed framework with 7 latest methods, including U-Net [24], Attention U-Net [22], R2U-Net [1],

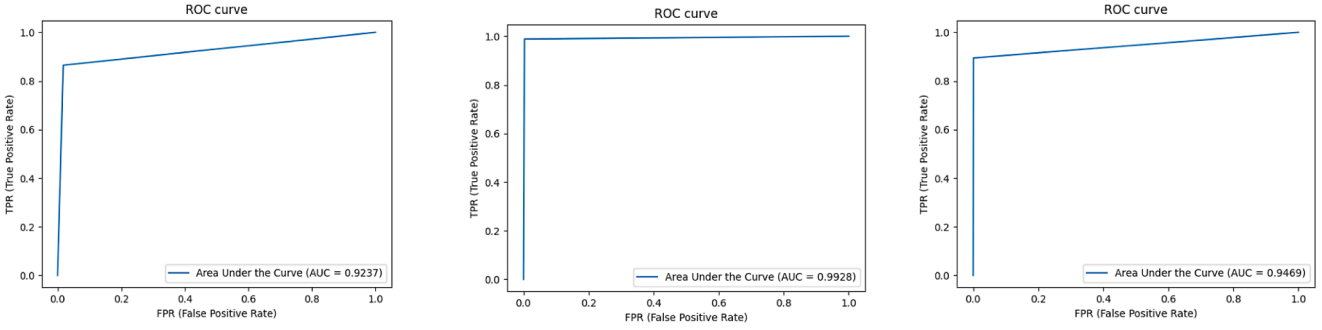
[3] https://challenge2018.isic-archive.com.
[4] https://www.kaggle.com/kmader/finding-lungs-in-ct-data.
[5] http://atriaseg2018.cardiacatlas.org.

Table 2
Comparison results on the Lung data set.

Method\ Metric	Dice	HD95	SP	SE	ACC	PC	JS
U-Net [24]	0.944 ± 0.146	6.002 ± 1.481	0.997 ± 0.003	0.946 ± 0.114	0.992 ± 0.004	0.949 ± 0.160	0.915 ± 0.162
Attention U-Net [22]	0.945 ± 0.151	5.890 ± 2.076	**0.998 ± 0.002**	0.933 ± 0.134	0.991 ± 0.009	0.963 ± 0.157	0.919 ± 0.163
R2U-Net [1]	0.950 ± 0.142	6.036 ± 1.833	0.995 ± 0.003	0.969 ± 0.096	0.993 ± 0.003	0.940 ± 0.157	0.926 ± 0.158
Attention R2U-Net [1]	0.951 ± 0.133	6.012 ± 1.653	0.995 ± 0.004	0.972 ± 0.064	0.993 ± 0.003	0.945 ± 0.153	0.926 ± 0.152
MSF-ACSA [7]	0.954 ± 0.127	6.971 ± 3.075	0.993 ± 0.008	0.961 ± 0.110	0.991 ± 0.006	0.951 ± 0.133	0.929 ± 0.139
BCDU-Net[2]	0.961 ± 0.125	5.136 ± 1.654	0.997 ± 0.003	0.969 ± 0.092	**0.995 ± 0.003**	**0.963 ± 0.131**	0.943 ± 0.143
U2-Net [23]	0.959 ± 0.129	5.193 ± 1.689	0.997 ± 0.002	0.971 ± 0.076	**0.995 ± 0.003**	0.958 ± 0.144	0.940 ± 0.145
Ours	**0.966 ± 0.115**	**4.947 ± 1.577**	0.997 ± 0.003	**0.983 ± 0.047**	**0.995 ± 0.003**	0.960 ± 0.131	**0.948 ± 0.131**

Table 3
Comparison results on the LA data set.

Method\ Metric	Dice	HD95	SP	SE	ACC	PC	JS
U-Net [24]	0.827 ± 0.180	3.724 ± 1.071	0.998 ± 0.002	**0.887 ± 0.163**	0.997 ± 0.002	0.802 ± 0.203	0.736 ± 0.205
Attention U-Net [22]	0.839 ± 0.168	3.662 ± 1.062	0.998 ± 0.002	0.881 ± 0.167	0.997 ± 0.002	0.828 ± 0.183	0.750 ± 0.195
R2U-Net [1]	0.846 ± 0.173	3.511 ± 0.934	**0.999 ± 0.001**	0.842 ± 0.183	0.997 ± 0.002	0.876 ± 0.175	0.762 ± 0.192
Attention R2U-Net [1]	0.847 ± 0.168	3.547 ± 0.982	**0.999 ± 0.001**	0.853 ± 0.185	0.997 ± 0.002	0.868 ± 0.169	0.762 ± 0.192
MSF-ACSA [7]	0.856 ± 0.143	3.492 ± 1.006	**0.999 ± 0.001**	0.878 ± 0.154	0.997 ± 0.002	0.857 ± 0.153	0.770 ± 0.172
BCDU-Net [2]	0.830 ± 0.186	3.636 ± 1.094	0.999 ± 0.002	0.823 ± 0.195	0.997 ± 0.003	0.873 ± 0.189	0.741 ± 0.209
U2-Net [23]	0.859 ± 0.161	3.337 ± 0.887	**0.999 ± 0.001**	0.850 ± 0.177	**0.998 ± 0.001**	**0.890 ± 0.161**	0.778 ± 0.185
Ours	**0.865 ± 0.149**	**3.323 ± 0.892**	**0.999 ± 0.001**	0.870 ± 0.157	**0.998 ± 0.001**	0.883 ± 0.154	**0.785 ± 0.180**



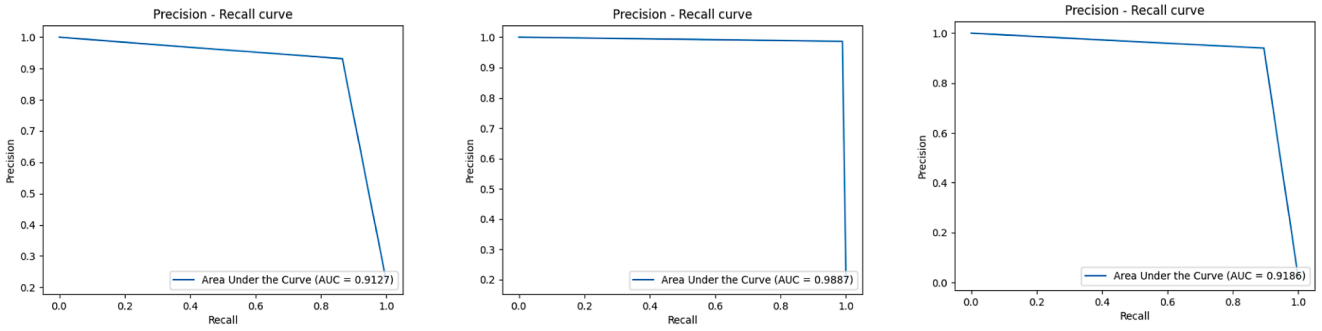
(a) ISIC (b) Lung segmentation (c) LA Segmentation

Fig. 4. ROC diagrams of the proposed method for three data sets.

Attention R2U-Net [1], MSF-ACSA [7], U2-Net [23], and BCDU-Net [2]. Please note that all results of the above methods are obtained by running the source code ourselves or by the author's pre-calculation. The source codes of the comparison methods for training and testing are provided in [22,1,23,2].

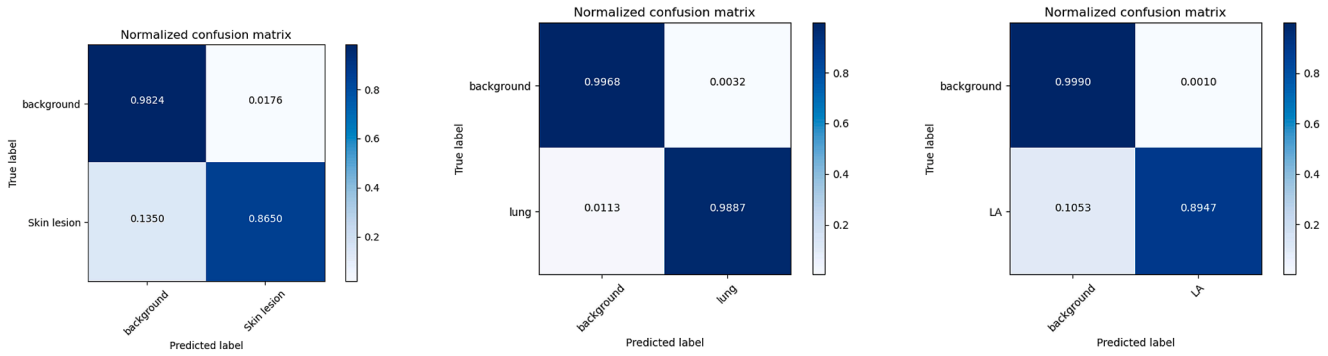
4.3.1. *Quantitative comparison*

To report the average performance and stability of all methods, each image must be tested separately. The experimental results on the three data sets are reported in Tables 1–3. In general, our method has achieved the best performance in two key indicators: Dice and HD95 with a big gap and other indicators have also achieved at least suboptimal performance. Specifically, for the segmentation of skin lesion, as shown in



(a) ISIC (b) Lung segmentation (c) LA Segmentation

Fig. 5. Precision-recall curves of the proposed method for three data sets.



(a) ISIC (b) Lung segmentation (c) LA Segmentation

Fig. 6. Confusion matrix presentations of the proposed method for three data sets.

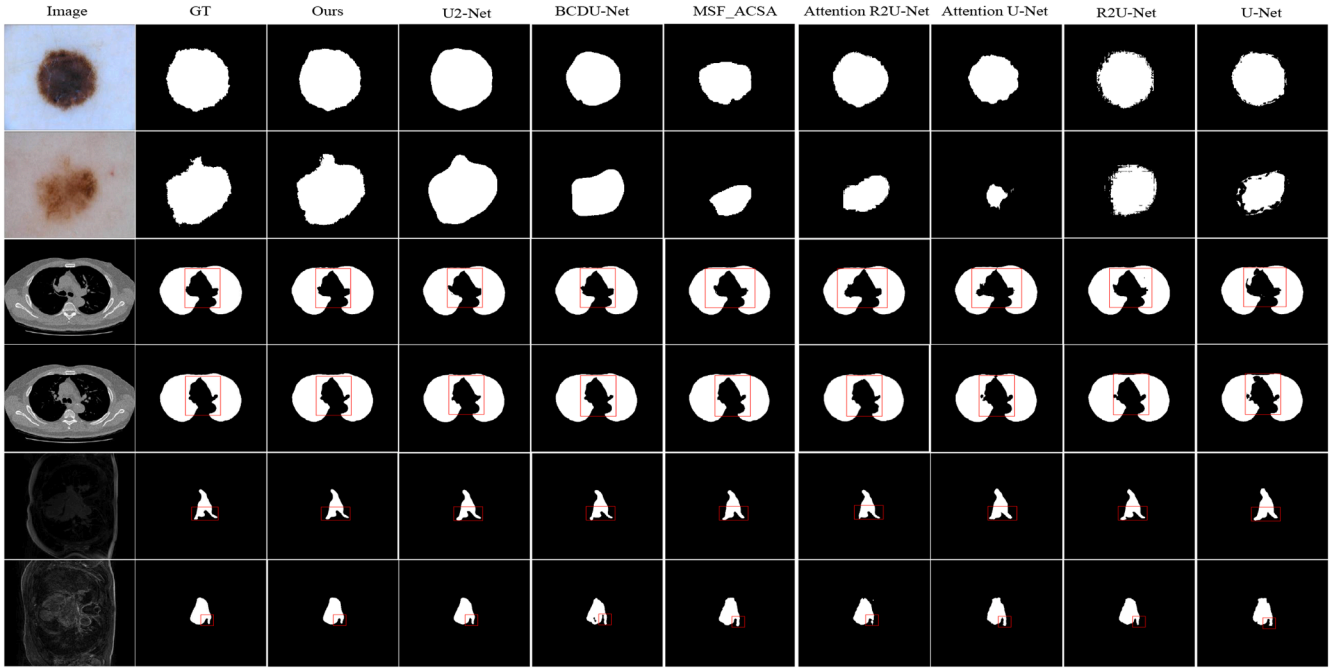


Fig. 7. Qualitative comparison with seven existing representative methods. The first two lines of the input data are skin disease cases from ISIC-2018, the middle two lines are from the Lung data set, and the last two lines are from the LA data set. Visual segmentation results are listed in each column by using different methods. It is easy to see that the results of the proposed method are closest to ground truth, and the visual effect is the best in both global positioning and local details, especially boundary details.

Table 1, the performance of our method greatly exceeds that of other methods on Dice and HD95, only slightly worse on SP and PC. For lung segmentation, as shown in Table 2, our framework is still only slightly worse in SP and PC too, and the others are optimal. For the segmentation of the left ventricle, as shown in Fig. 3, the conclusion is similar to the first two cases. Dice indicates the coincidence degree between the prediction and GT. The larger the value, the higher the coincidence degree. Hd95 implies a distance between the prediction and GT boundary. The smaller the distance, the closer the two are, the better the prediction result will be. In general, the performance of our method is the best in these two indicators, but it is deficient in both SP and PC.

As shown in Table 1, Our results are significantly better than all other methods in Dice, ACC and JS, and the HD95 is obviously optimal/minimum, which shows that the coincidence degree between the segmented region inferred by our model and the manually labeled mask is the highest, and the edge similarity is also the highest. From the standard deviation, our result is also the most stable. The conclusion in Table 2 is similar to that in Table 1. Only BCDU-Net and u2net have the same performance as our method. Our performance is almost all-around. Only our PC index is worse than BCDU-Net, which is enough to show the superiority of our proposed model. Surprisingly, the performance of BCDU-Net in Table 3 is poor. U2-Net and our method are almost always the best two methods, followed by the MSF-ACSA model. U2-Net is only slightly better than our method in the PC index.

To show the overall performance of the network on three data sets, ROC curves, precision recall curves and confusion matrices are shown in Figs. 4–6 respectively. Here, we flatten all test images and corresponding labels into 1D arrays respectively for overall calculation. Unlike the data in Tables 1–3, each image is evaluated separately, and then the mean and standard deviation are obtained. For the above two kinds of curves, the shape is not very important, but the key lies in the area formed with the X-axis. To show the performance more intuitively, we list the three normalized confusion matrices on the three data sets.

Table 4
Ablation analysis of our method on ISIC-2018.

Method\Metric	Dice	HD95
U2-Net	0.8804 ± 0.1449	4.3707 ± 1.9913
U2-Net+MSI	0.8810 ± 0.1447	4.3705 ± 1.9912
U2-Net+MSI+SDM	0.8833 ± 0.1385	4.3688 ± 1.9856
U2-Net+MSI+SDM+RM	**0.8871 ± 0.1327**	**4.3604 ± 1.9739**

MSI: Multi-Scale Inputs. RM: Refined Module. SDM: Signed Distance Map regression branch. The best results are marked in bold.

4.3.2. *Visual comparison*

In Fig. 7, we show some visualization results to evaluate the effectiveness of our method in object segmentation. From the segmentation results, benefiting from the enhancement of edge features, we can see that our method can not only accurately locate the whole segmentation region, but also highlight the sharp and boundaries of the objects. For instance, for the first and second samples with very smooth edges and relatively sharp, most comparative methods can only locate those approximate positions. However, based on the additional SDM branch and the subsequently refined networks, our method can well capture the edge features and achieve better accuracy than other methods. For the third and fourth samples, almost all methods have achieved good results in the outer boundary segmentation, but the inner gap is large. As a whole,the edge information of the segmented object is not prominent, and our results are still the closest to the ground truth. The last two lines show the evaluation results of two samples from the LA test set, where u2net, BCDU-Net, and MSF_ACSA can achieve better performance than the comparison methods. After careful comparison, our results are still closest to the ground truth.

4.4. *Ablation study*

To verify the feasibility and effectiveness of our proposed framework, we conduct an ablation study on the ISIC-2018 data set. Two different variants in the last two rows of the proposed model are tested on the same dataset, one of which has a refined module and the other does not. The hyper-parameter settings of the two models are the same. In addition, we also add the first two rows as a comparison to verify the impact of multi-scale input. We first train these models to converge, and then evaluate them, as shown in Table 4.

From Table 4, we can see that the SDM branch and the refined module can improve overall performances indeed. In addition, the multi-scale input can slightly improve the overall performance.

5. Conclusion

This paper presented a region-to-boundary and coarse-to-fine deep learning framework for medical image segmentation to ensure overall segmentation performance and highlight boundary details. First, four identical samples of different sizes are input into different layers of the encoder of the proposed DL framework. Multi-scale features are extracted from each layer of the decoder for deep supervision. In addition, an SDM regression branch is generated at the end of the decoder to obtain the edge probability map of the target, which can well guide the subsequent network to segment the boundaries. Finally, we obtain the segmentation results by fusing the edge features and overall features. With the help of SDM, we can enhance edge features and improve boundary segmentation performance. Experiments are performed to compare the proposed framework with the latest methods on three public datasets, and demonstrate the advantages of our framework in medical image segmentation.

CRediT authorship contribution statement

Xiaowei Liu: Conceptualization, Methodology, Software, Writing - original draft. **Lei Yang:** Data curation, Investigation. **Jianguo Chen:** Writing - review & editing, Funding acquisition. **Siyang Yu:** Software, Validation. **Keqin Li:** Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Thank the reviewers for their hard work and give many constructive suggestions for improving the quality of this manuscript. This research is supported by the National Key R&D Program of China under Grant 2018YFB1003401, and the National Natural Science Foundation of China under Grant 6200211.

References

[1] M.Z. Alom, M. Hasan, C. Yakopcic, T.M. Taha, V.K. Asari, Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation; 2018. arXiv preprint arXiv:1802.06955.
[2] M. Asadi-Aghbolaghi, R. Azad, M. Fathy, S. Escalera, Multi-level context gating of embedded collective knowledge for medical image segmentation; 2020. arXiv preprint arXiv:2003.05056.
[3] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional convlstm u-net with densley connected convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2019, pp. 406–415.
[4] G. Borgefors, Distance transformations in digital images, Computer Vision, Graphics, and Image Processing 34 (1986) 344–371.
[5] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3d u-net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2016, pp. 424–432.
[6] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[7] C. Gao, H. Ye, F. Cao, C. Wen, Q. Zhang, F. Zhang, Multiscale fused network with additive channel–spatial attention for image segmentation, Knowledge-Based Systems 214 (2021), 106754.
[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[9] S. Guan, A. Khan, S. Sikdar, P. Chitnis, Fully dense unet for 2d sparse photoacoustic tomography artifact removal, IEEE Journal of Biomedical and Health Informatics (2019).
[10] A. Hatamizadeh, D. Terzopoulos, A. Myronenko, End-to-end boundary aware networks for medical image segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2019, pp. 187–194.
[11] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[12] S. Hu, J. Zhang, Y. Xia, Boundary-aware network for kidney tumor segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2020, pp. 189–198.
[13] N. Ibtehaz, M.S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Networks 121 (2020) 74–87.
[14] V. Iglovikov, A. Shvets, Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation, 2018, arXiv preprint arXiv:1801.05746.
[15] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, nnu-net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2021) 203–211.
[16] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.
[17] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, I.B. Ayed, Boundary loss for highly unbalanced segmentation, Medical Image Analysis 67 (2021), 101851.
[18] X. Liu, K. Li, K. Li, Attentive semantic and perceptual faces completion using self-attention generative adversarial networks, Neural Processing Letters 51 (2020) 211–229.
[19] R. McKinley, R. Meier, R. Wiest, Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 456–465.
[20] A. Myronenko, 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 311–320.
[21] A. Myronenko, A. Hatamizadeh, 3d kidneys and kidney tumor semantic segmentation using boundary-aware networks, 2019, arXiv preprint arXiv:1909.06684.

[22] O. Oktay, J. Schlemper, L.L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, et al., Attention u-net: Learning where to look for the pancreas, 2018, arXiv preprint arXiv:1804.03999.
[23] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O.R. Zaiane, M. Jagersand, U2-net: Going deeper with nested u-structure for salient object detection, Pattern Recognition 106 (2020), 107404.
[24] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-assisted Intervention, Springer, 2015, pp. 234–241.
[25] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[26] X. Wu, H. Chen, Y. Huang, H. Guo, T. Qiu, L. Wang, Center-sensitive and boundary-aware tooth instance segmentation and classification from cone-beam ct, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 939–942.
[27] X. Xiao, S. Lian, Z. Luo, S. Li, Weighted res-unet for high-quality retina vessel segmentation, in: 2018 9th International Conference on Information Technology in Medicine and Education (ITME), IEEE, 2018, pp. 327–331.
[28] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[29] B. Zhou, R. Crawford, B. Dogdas, G. Goldmacher, A. Chen, A progressively-trained scale-invariant and boundary-aware deep neural network for the automatic 3d segmentation of lung lesions, in: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019, pp. 1–10.