Mathematical Biosciences and Engineering

MBE, 20(1): 365–382.
DOI: 10.3934/mbe.2023017
Received: 27 April 2022
Revised: 18 September 2022
Accepted: 28 September 2022
Published: 08 October 2022

http://www.aimspress.com/journal/MBE

Research article

DSCA-Net: A depthwise separable convolutional neural network with attention mechanism for medical image segmentation

Tong Shan[1], Jiayong Yan[2,3,*], Xiaoyao Cui[3] and Lijian Xie[4,*]

[1] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[2] School of Medical Instruments, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China
[3] Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, Suzhou 215163, China
[4] Children's Hospital of Shanghai, Shanghai 200062, China

* **Correspondence:** Email: yanjy@sumhs.edu.cn, naijileix@aliyun.com.

Abstract: Accurate segmentation is a basic and crucial step for medical image processing and analysis. In the last few years, U-Net, and its variants, have become widely adopted models in medical image segmentation tasks. However, the multiple training parameters of these models determines high computation complexity, which is impractical for further applications. In this paper, by introducing depthwise separable convolution and attention mechanism into U-shaped architecture, we propose a novel lightweight neural network (DSCA-Net) for medical image segmentation. Three attention modules are created to improve its segmentation performance. Firstly, Pooling Attention (PA) module is utilized to reduce the loss of consecutive down-sampling operations. Secondly, for capturing critical context information, based on attention mechanism and convolution operation, we propose Context Attention (CA) module instead of concatenation operations. Finally, Multiscale Edge Attention (MEA) module is used to emphasize multi-level representative scale edge features for final prediction. The number of parameters in our network is 2.2 M, which is 71.6% less than U-Net. Experiment results across four public datasets show the potential and the dice coefficients are improved by 5.49% for ISIC 2018, 4.28% for thyroid, 1.61% for lung and 9.31% for nuclei compared with U-Net.

Keywords: medical image segmentation; lightweight neural network; attention mechanism

1. Introduction

Accurate medical images segmentation is basic and crucial for medical image processing and analysis [1,2]. Generally, the targets on medical images are segmented by sketching the outline manually, but this is time-consuming and requires professional knowledge of physicians. Lots of morphology-based automatic segmentation methods have been proposed in the past, including edge detection, area detection, and template matching [3]. However, it is difficult to design specifically and easily deformable models for various segmentation tasks [4]. The significant variations in the scale and shape of segmented targets add to the difficulty of segmentation tasks [5].

With the great development of deep learning, deep convolutional neural networks (DCNNs) have achieved excellent performance in medical image segmentation filed [3,6,7]. Compared with traditional methods, DCNNs can automatically extract features and show higher accuracy and robustness. Many structures [8–10] were founded on Fully Convolutional Network (FCN), and U-Net [10] with its variants, have been widely implemented to many tasks, such as skin lesions [11,12], thyroid gland [13,14], lung [7,15], nuclei [16–19], etc. U-Net adopts the encoder-decoder structure. The encoder captures feature information using continuous stacked convolutional layers and the decoder is sited to recover the categories of each pixel. At the same time, multiple skip connections are also applied to spread feature information for final segmentation. The variants, such as MultiResUNet [12], UNet++ [20], rethought the U-Net architecture and achieved better performance in segmentation tasks. Nevertheless, there are still some shortcomings of U-Net and its variants. First, the pooling operation may lose some important features which are conductive to improve the segmentation accuracy. Second, these methods couldn't dynamically adjust to the variation of features, such as shape and size. Third，continuous stacked convolution layers deepen the network architecture and enhance the feature extraction capability of models, but a major critique of such models is their large parameter count [21].

The attention mechanism helps the network to draw what we need. By imitating the way humans allocate attention, the attention feature vectors or maps dynamically add the important weights to critical information and omit useless ones. Using the squeeze-and-excitation module, SE-Net [22], showed effectiveness of the inter-channel relationship. ECA module [23], inspired by SE-Net, generated feature weights by 1D convolution operation and effected on output. [24] adopted the attention mechanism for image classification based on RNN model and achieved good performance. [25] first applied attention mechanism to the field of NLP in machine translation tasks, and [26] proposed a self-attention mechanism. For medical image segmentation field, Attention U-Net [27] applied attention gates for capturing richer contextual information. CA-Net [11] proposed a comprehensive attention network to emphasize significant features in multi-scale feature maps. These methods take advantage of context information and achieve higher accuracy, but the parameters are relatively higher.

The emergence of depthwise separable convolution has shown great efficiency and reduced training parameters over regular convolution [28–30]. It separates the standard convolution operation into two layers: depthwise convolution and pointwise convolution. Each input channel is first convoluted spatially, and the pointwise convolution subsequently processes the channels into a new channel dimensional space, subsequently. In MobileNets architecture [29], depthwise separable convolution was employed to build lightweight networks and embedded in mobile visual applications.

DeepLabV3+ [31] applied it to the ASPP module, which achieved a faster and more powerful network for semantic image segmentation. X-Net [32] adopted it to scale the network size down and performed well. MobileNetV3-UNet [33] created a lightweight encoder and decoder architecture based on depthwise separable convolution, which achieved high accuracy on medical image segmentation tasks.

Combining the advantages of attention mechanism and depthwise separatable convolution into U-shaped architecture, a lightweight DSCA-Net is proposed in this paper for medical image segmentation. Three novel attention modules are proposed and integrated into the encoder and decoder of U-Net, separately. The chief contributions of our work are summarized as follows:

1) A Pooling Attention module is proposed to reduce the feature loss caused by down-sampling.

2) A Context Attention module is designed to exploit the concatenation feature maps from the encoder and decoder, which combines the spatial and channel attention mechanisms to focus on useful position features.

3) To make better use of multi-scale information from different level stages of the decoder, a Multiscale Edge Attention module is proposed to deal with combined features for the final prediction.

4) We integrate all proposed modules into DSCA-Net for medical image segmentation and all convolution operations are implemented by depthwise separable convolution. The proposed network was evaluated on four public datasets and the experimental results reveal that our proposed network outperforms previous state-of-the-art frameworks.

The remainder of our paper is structured below. Section 2 goes over detailed information of proposed DSCA-Net architecture, and Section 3 describes the experimental settings and results. Finally, some discussions and conclusions are given in Sections 4 and 5.

2. Materials and methods

2.1. DSCA-Net

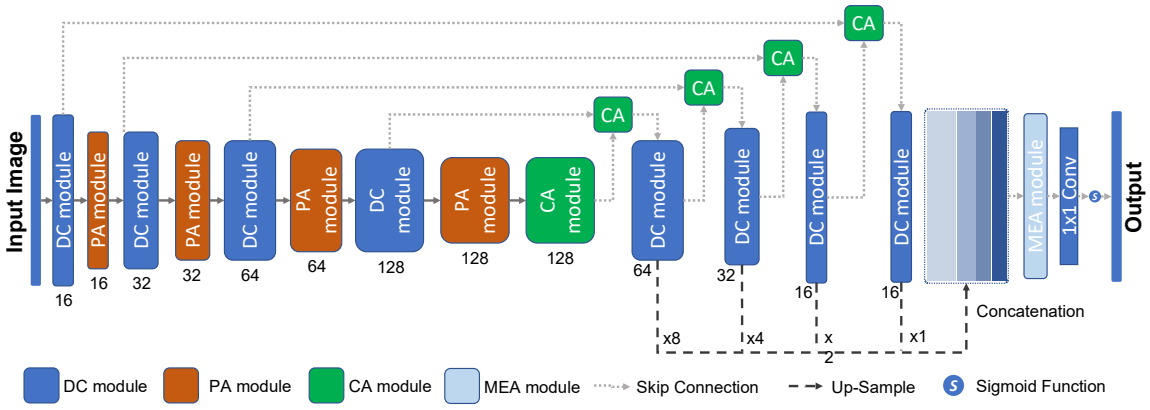

Figure 1. The overall architecture of DSCA-Net.

By combining attention mechanism and depthwise separable convolution with the architecture of U-Net, we propose DSCA-Net, which is shown in Figure 1. The network is composed of encoding part, decoding part, and multiscale edge part. Firstly, we replace the stacked 3×3 convolution layers of U-Net with DC module. The depth of encoder is 128, which enables our proposed model better extracting abundant features while reducing parameter amount. Secondly, to reduce the feature loss,

PA module is embedded in place of maximum pooling layer, which has almost no effect on the number of parameters. Then, long-range skip connections are utilized to transfer feature maps from encoder to symmetrical decoder stage after passing through CA module, which fuses and recalibrates the context information at five different resolution levels. Finally, MEA module reemphasizes the salient scale information from concatenating multiscale feature maps, which enable the last CNN layer to be aware of segmenting target edge.

2.2. Dense convolution module

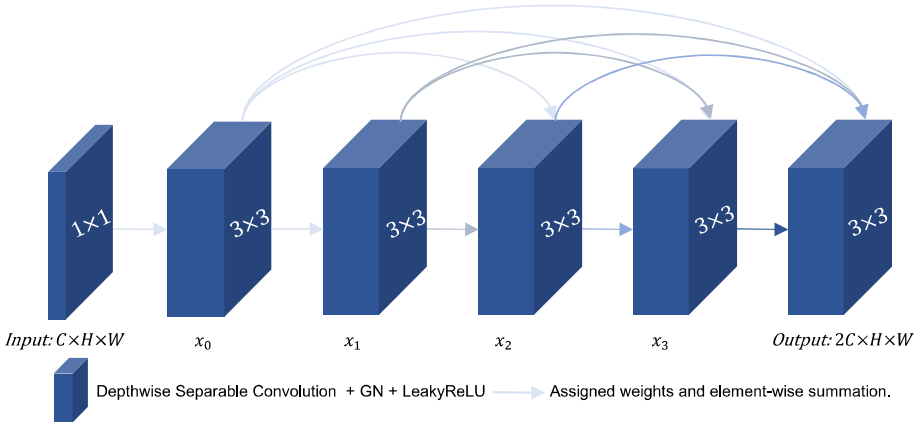

Figure 2. Dense convolution module.

Recent studies show that extending the network depth leads to better segmentation performance [28,34]. Based on depthwise separable convolution operation [28] and DenseNet [35], the dense convolution module is proposed. We utilize it in encoder to extract high-dimensional feature information and recover segmented target details in decoder. As shown in Figure 2, every depthwise separable convolution layer is followed by one group normalization [36] and LeakyReLU [37], which improves nonlinear expression capability of model. For convenience, we assume the input as $x_{input} \in \mathbb{R}^{C\times H\times W}$. C, H, W denote channel, height, and weight, respectively. At beginning, one 1×1 convolution layer $F_{1\times1}^{conv}$ expands x_{input} channel numbers 2 times. Then, multiple residual connections from former layers are summed to all subsequent layer with two continuous 3×3 convolution layers $F_{3\times3}^{conv}$. The elementwise summation operations are used for fusing extracted information without adding parameters. DC module is described by the following equation:

$$x_0 = F_{1\times1}^{conv}(x_{input}) \tag{1}$$

$$x_i = F_{3\times3}^{conv}(SUM(x_0; x_1; x_2; \ldots; x_{i-1})) \tag{2}$$

where $x_0 \in \mathbb{R}^{C\times H\times W}$ denotes the input feature map and $x_i \in \mathbb{R}^{C\times H\times W}$ represents the feature maps in layer i.

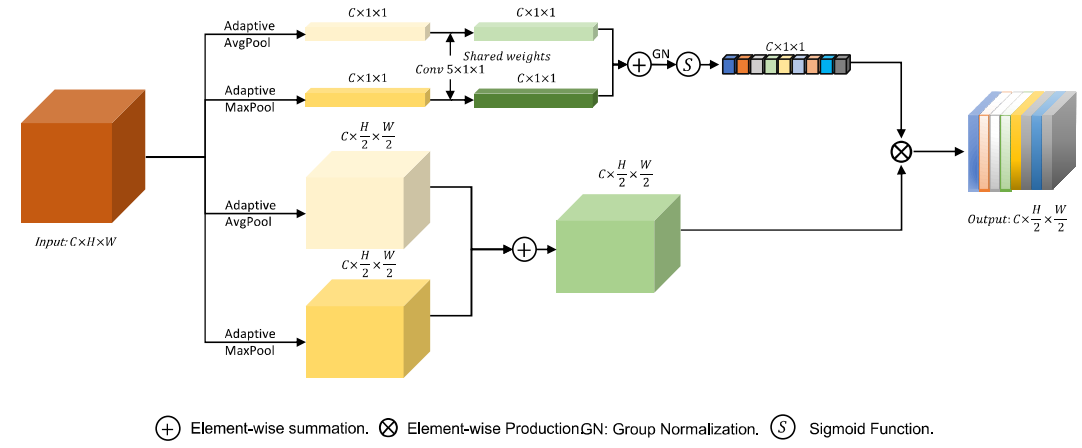

Figure 3. Pooling attention module.

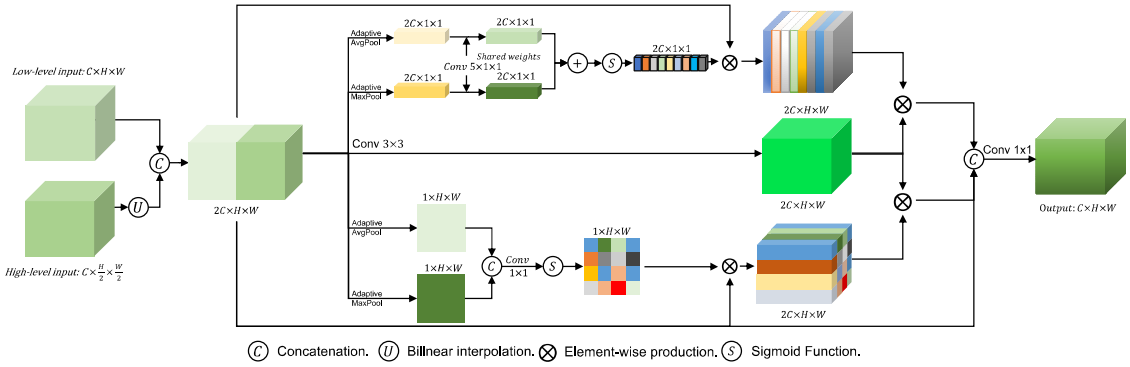

Figure 4. Context attention module.

2.3. Pooling attention module

Consecutive pooling operation in encoder enlarges the reception of convolution operation but lose certain features. Therefore, we rethink SE-Net [22] and ECA-Net [23], and propose PA module to replace the original pooling layer, as shown in Figure 3. PA module mainly takes in a two-branch structure. One branch tries to obtain an attention channel feature vector and the other rescales height and width of feature maps. First, a 1D convolution $F_{5\times1\times1}^{conv}$ with shared kernel weights of 5 is used to extract more abundant feature information after adaptive maximum pooling $P_{1\times1}^{max}$ and average pooling $P_{1\times1}^{avg}$ layers. Then, the vector V_{sum} is summed element-by-element and activated by $Sigmoid$ function. Finally, the output y_{out} is multiplied by rescaled feature maps M_{scaled}. PA module can be expressed as follows:

$$V_{sum} = Sigmoid(GN(F_{5\times1\times1}^{conv}(P_{1\times1}^{max}(x_{input})) \oplus F_{5\times1\times1}^{conv}(P_{1\times1}^{avg}(x_{input})))) \tag{3}$$

$$M_{scaled} = P_{\frac{h}{2}\times\frac{w}{2}}^{max}(x_{input}) \oplus P_{\frac{h}{2}\times\frac{w}{2}}^{avg}(x_{input}) \tag{4}$$

$$y_{out} = V_{sum} \otimes M_{scaled} \tag{5}$$

where $x_{input} \in \mathbb{R}^{C\times H\times W}$ denotes input feature maps, $\oplus$ and $\otimes$ denote element-wise summation and element-wise production, respectively.

2.4. Context attention module

In the process of context information extraction, simple concatenation of U-Net is not sufficient to gradually restore needed information. Drawing lessons from dynamic weight similarity calculation, we propose CA module to fuse context information, as shown in Figure 4. $x_{low} \in \mathbb{R}^{C\times H\times W}$ and $x_{high} \in \mathbb{R}^{C\times\frac{H}{2}\times\frac{W}{2}}$ represent feature maps from encoder and decoder, respectively. At first, we obtain $x_{input} \in \mathbb{R}^{2C\times H\times W}$ via concatenating x_{low} and x_{high} from upper decoder layer. Then, to capture detailed context information, CA module adopts a three-branch structure, including a spatial attention branch, a channel attention branch, and a convolution branch, which has the same dimensions of x_{input} and y_{out}. The learned feature maps from spatial attention branch $x_{spatial} \in \mathbb{R}^{2C\times H\times W}$ and channel attention branch $x_{channel} \in \mathbb{R}^{2C\times H\times W}$ multiply convolutional feature maps $x_{conv} \in \mathbb{R}^{2C\times H\times W}$, separately. Finally, feature maps are concatenated and one 1×1 convolution $F_{1\times1}^{conv}$ reconstructs $y_{out} \in \mathbb{R}^{C\times H\times W}$. The relevant formula can be stated as follows:

$$x_{input} = Cat\left[x_{low}; U\left(x_{high}\right)\right] \tag{6}$$

$$x_{channel} = x_{input}\otimes(Sigmoid(F_{5\times1\times1}^{conv}(P_{1\times1}^{max}\left(x_{input}\right)) \oplus F_{5\times1\times1}^{conv}(P_{1\times1}^{avg}\left(x_{input}\right)))) \tag{7}$$

$$x_{spatial} = x_{input}\otimes(Sigmoid(F_{1\times1}^{conv}(Cat[P_{H\times W}^{max}\left(x_{input}\right);\ P_{H\times W}^{avg}\left(x_{input}\right)]))) \tag{8}$$

$$y_{out} = F_{1\times1}^{conv}\left(Cat\left[x_{spatial}\otimes x_{conv}; x_{channel}\otimes x_{conv}; x_{input}\right]\right) \tag{9}$$

where $Cat[\cdot]$ denotes the concatenation operation along with channel dimension. $P_{H\times W}^{avg}$ denotes adaptive average pooling operation and $P_{H\times W}^{max}$ denotes adaptive maximum pooling operation. For the process of bottom feature information, we use a deformable CA module to capture context information with single input.

2.5. Multiscale edge attention module

U-Net uses decoder to restore the categories of each pixel. However, segmented objects with large variant scales and blurred edges increase the difficulty of accurate segmentation. The pixel position of target edge in feature maps of union scales from decoder is slightly different and the high-level feature maps in decoder contain more sufficient edge information. Learning the scale-dynamic weights of all fused feature map pixels for calibrating object edge is desirable. To utilize multiscale feature maps, we propose MEA module, as shown in Figure 5. First, we use bilinear up-sampling layers of different

scale factors ($s = 1, 2, 4, 8$) to unify feature map scale obtained from decoder to the final output size and concatenate them. Then, for learning scale-dynamic features, one 1×1 convolution and $Sigmoid$ function generate calibrated weights and multiplied original input to obtain $y_{out} \in \mathbb{R}^{C \times H \times W}$. MEA module can be described as follows:

$$x_{input} = Cat[s(x_1); s(x_2); \ldots s(x_i)] \tag{10}$$

$$y_{out} = x_{input} \otimes GN\left(Sigmoid\left(x_{input}\right)\right) \tag{11}$$

where $s(\cdot)$ denotes resampled function with different scale factors. x_{input} denotes concatenated feature map and GN denotes Group Normalization.

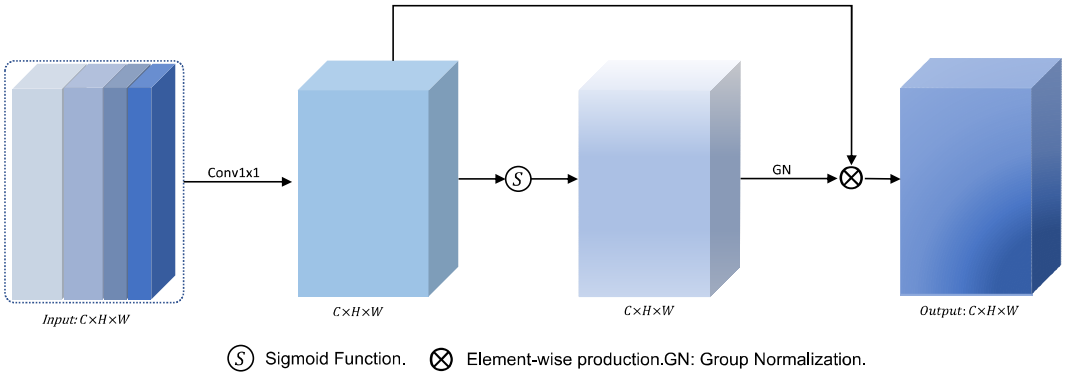

Figure 5. Multiscale edge attention module.

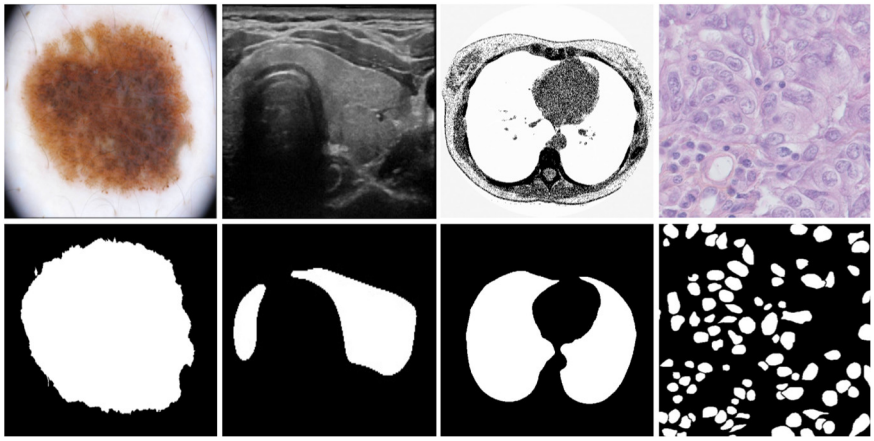

Figure 6. Samples of four datasets.

3. Experiments and results

To assess our proposed network, we validated DSCA-Net and compared with other state-of-the-art methods on four public datasets: ISIC 2018 dataset [5,38], thyroid gland segmentation dataset [39], lung segmentation (LUNA) dataset, and nuclei segmentation (TNBC) dataset [17]. Each dataset poses its separate challenge, and the corresponding samples are shown in Figure 6. On each task, we compared

the results with state-of-the-art networks and implemented ablation studies to demonstrate effectiveness of modules, which will be discussed in Sections 3.3–3.6.

3.1. Experiment setup

During the experimental period, all models in this paper were achieved based on Pytorch and the experimental planform was supported by Linux 18.04 operating system, which was equipped with Intel Xeon CPU @2.30 GHz and 27GB RAM. The GPU was 16 GB Nvidia Tesla P100-PCIE. The Adam optimizer [40] was used with learning rate 10^{-4}, and weight decay 10^{-8}. The dynamic learning rate was decayed by 0.5 every 100 epochs. We utilized the Soft Dice loss for model training and kept optimal result upon validation dataset. Quantitative results were obtained in test.

To maximize the use of GPU, the batch sizes are set to 8, 4, 12, and 2 for ISIC, thyroid gland, LUNA, and TNBC datasets, respectively. For better fitting data, the number of iterations for TNBC dataset is 500, and 300 for others. The training process stops automatically after the maximum epoch. We utilized Fivefold cross-validation for result to assess the stability and effectiveness of DSCA-Net. Every input image was normalized from $[0, 255]$ to $[0, 1]$. During model training, random rotation and flipping of the angle in $(-\frac{\pi}{9}, \frac{\pi}{9})$ with the probability of 0.5 were applied for data augmentation.

3.2. Evaluation metrics

In this paper, Dice coefficient (Dice), Intersection over Union (IoU), accuracy (Acc), specificity (Spec), sensitivity (Sens) and average symmetric surface distance (ASSD) are used as evaluation metrics. The formula for all metrics can be expressed as follows:

$$Dice = \frac{2*TP}{2TP+FN+FP} \tag{12}$$

$$IoU = \frac{TP}{TP+FN+FP} \tag{13}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{14}$$

$$Specific = \frac{TN}{TN+FP} \tag{15}$$

$$Sensitivity = \frac{TP}{TP+FN} \tag{16}$$

where TP, TN, FP, FN represent the predicted pixel numbers of true positive, true negative, false positive, and false negative, respectively. Assuming S_a and S_b are the set of border points from prediction result and corresponding label, individually, $ASSD$ is defined as:

$$ASSD = \frac{\left(\sum_{a\in S_a} d(a,S_b)+\sum_{b\in S_b} d(b,S_a)\right)}{|S_a|+|S_b|} \tag{17}$$

where $d(v, S_a) = min_{x \in S_a}(|v - x|)$ represents the shortest Pythagorean distance between point v and S_a.

3.3. Skin lesion segmentation

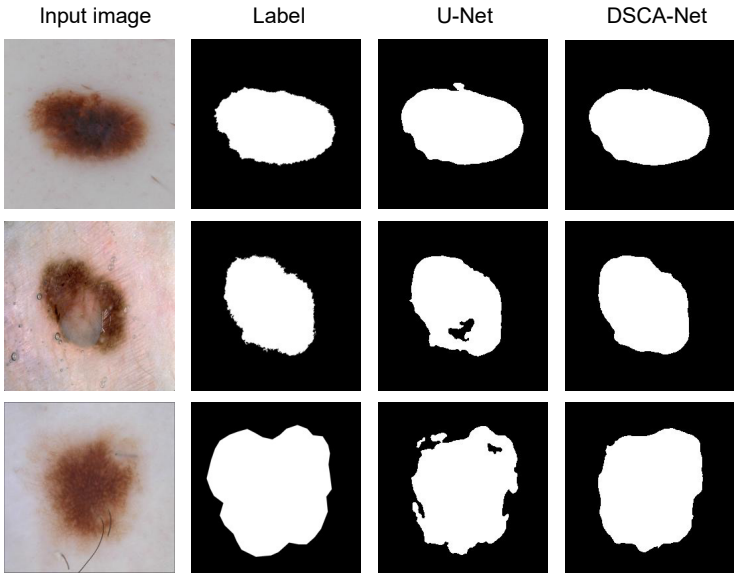

Figure 7. Visualization results of skin lesion segmentation dataset.

Table 1. Comparisons of segmentation performance and number of parameters between DSCA-Net and other networks on skin lesion segmentation.

Methods	Dice	Acc	Params
U-Net [10]	0.8739	0.8777	7.76 M
Attention-UNet [27]	0.8846	0.8846	34.88 M
RefineNet [41]	0.9155	0.9155	46.3 M
EOC-Net [42]	0.8611	0.8401	-
CA-Net [11]	0.9208	0.9268	2.8 M
DeepLabV3+ [43]	0.9221	0.9179	54.7 M
Mobi leNetV3-UNet [33]	0.9098	0.9479	8.3 M
IBA-U-Net [44]	-	0.9440	13.91 M
Ours	0.9282	0.9532	2.2 M

The skin lesion segmentation dataset has 2594 images and their corresponding label in 2018 [5,38]. We randomly divided the dataset by the ratio of 7:2:1 into 1815, 261, and 520 used for training, validation, and testing, respectively. The original size of images in dataset varies from 720×540 to 6708×4439. To facilitate the training process of our proposed network, all images and corresponding masks were cropped to 256×256.

Some skin lesion segmentation samples of our proposed network and U-Net are shown in Figure 7. U-Net performs unsatisfactorily compared with DSCA-Net in regular skin lesion segmentation images. When the skin lesion has a similar color to surroundings or occluded by hair and tissue fluid, U-Net gets error segmentation results. The more blurred boundary of skin lesion, the more incorrect segmentation is obtained by U-Net. Comparatively, DSCA-Net performs better.

To fully confirm the validity of our method, we compared DSCA-Net with U-Net [10], Attention U-Net [27], RefineNet [41], EOCNet [42], CA-Net [11], DeepLabv3+ [31], MobileNetV3-UNet [33] and IBA-U-Net [44] on this dataset. The results are listed in Table 1. Our proposed model performs an Acc of 0.9532, 0.0755 higher than U-Net, 0.0053 higher than second-place method MobileNetV3-UNet. Although Dice is 0.0002 less than DeepLabV3+, the difference is not significant. Our model has 1/3.53, 1/15.85, 1/24.86, 1/1.27, 1/3.77 and 1/6.32 times fewer parameters than U-Net, attention U-Net,DeepLabV3+, CA-Net, MobileNetV3-UNet, and IBA-U-Net with better segmentation performance, respectively.

Table 2. Quantitative evaluation of ablation study on skin lesion segmentation dataset.

Methods	Dice	IoU	ASSD
Lightweight U-Net (Backbone)	0.8232	0.7164	1.6189
Backbone + DC	0.8612	0.7672	1.0833
Backbone + DC + PA	0.8941	0.8157	0.8364
Backbone + DC + PA + CA	0.9219	0.8711	0.6753
DSCA-Net	0.9282	0.8733	0.6318

Table 2 lists the comparison results. Lightweight U-Net were achieved by depthwise separable convolution instead of original stacked convolution layers of U-Net. DSCA-Net is the network adding all designed modules. The quantitative results show that our proposed modules strengthen the feature extraction ability. Every proposed module improves segmentation performance. At the same time, Backbone + DC + PA, and Backbone + DC + PA + CA shows better segmentation results than U-Net.

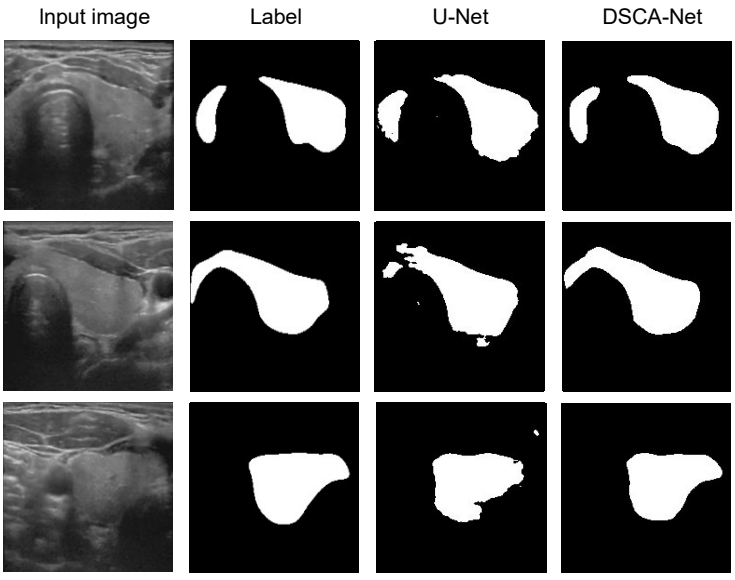

Figure 8. Visualization results of thyroid gland segmentation dataset.

3.4. Thyroid gland segmentation

The thyroid public dataset [39] was acquired by a GE Logiq E9 XDclear 2.0 system equipped with a GE ML6-15 ultrasound probe with Ascension driveBay electromagnetic tracking. It took from

healthy thyroid records and the volumes were taken straight from the ultrasound imaging instrument, which was recorded in DICOM format. The matching label, which was produced by a medical expert, contains the isthmus as part of the segmented region. To train our model, we split the volume into 3998 individual slices with label. We randomly used 2798 images for training, 400 images for validation and 800 for testing, with a ratio of 7:1:2. The shape of input was randomly cropped in 256×256.

Table 3. Comparisons of segmentation performance and number of parameters between DSCA-Net and other networks on thyroid gland segmentation.

Methods	Dice	Sens	Spec	Params
U-Net [10]	0.9332	0.9526	0.9169	7.76 M
SegNet [8]	0.8401	0.9811	0.8437	112.32 M
SUMNet [14]	0.9207	0.9830	0.8911	-
Attention-UNet [27]	0.9582	0.9801	0.9444	34.88 M
DSCA-Net	0.9727	0.9873	0.9921	2.2 M

Table 4. Quantitative evaluation of ablation study on thyroid gland segmentation dataset.

Methods	Dice	IoU	ASSD
Lightweight U-Net (Backbone)	0.8837	0.8079	1.0703
Backbone + DC	0.9017	0.8325	0.8331
Backbone + DC + PA	0.9113	0.8469	0.6557
Backbone + DC + PA + CA	0.9687	0.9422	0.1072
DSCA-Net	0.9727	0.9544	0.0953

Figure 8 presents several test segmenting results on thyroid gland dataset. The edge of thyroid gland and background information usually have some outliers and similarities in vision, but not relate to our interest. Observations show that U-Net under-segmented thyroid isthmus while DSCA-Net better.

We tested DSCA-Net against three methods: SegNet [8], SUMNet [14], and Attention-UNet [27]. Quantitative evaluation results present in Table 3. The Dice increases from 0.9332 to 0.9727 by 4.2%, the Sens increases from 0.9526 to 0.9873 by 3.6% and Spec increases from 0.9169 to 0.9921 by 8.2% compared with U-Net. Our model has 1/51.05 times fewer parameters than SegNet and performs better through evaluation metrics.

Additionally, Table 4 presents the quantitative analysis results of ablation study on thyroid segmentation. DSCA-Net scored the best performance in every metric. The Dice increased significantly after adding CA module, which indicates that CA module can efficiently extract context information for thyroid segmentation performance.

3.5. Lung segmentation

Lung segmentation requires segmenting the lung structure from a competition called Lung Nodule Analysis (LUNA). It contains 534 2D CT samples with corresponding label. The original resolution of images is 512×512, and we randomly cropped them into 256×256. Separately, 70, 10, and 20% of dataset is allocated for training, validation, and testing with corresponding number 374,

53, and 107.

From the visualization results shown in Figure 9, DSCA-Net performs better than U-Net in detailed edge processing. Affected by the noise of lung CT images, U-Net produces some erroneous segmented areas. DSCA-Net has a greater tolerance to noise than U-Net. We demonstrate the validation of our approach by achieving a promising improvement despite the relatively simple task.

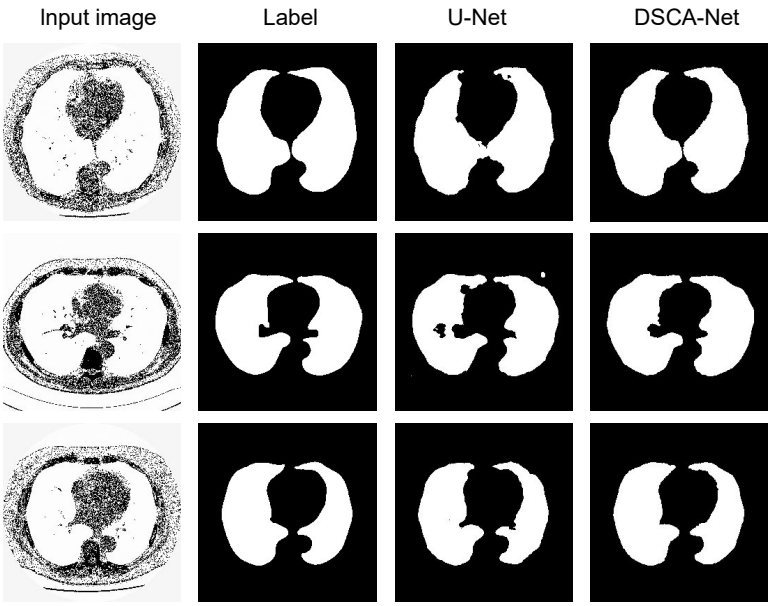

Figure 9. Visualization results of lung segmentation dataset.

Table 5. Comparisons of segmentation performance and number of parameters between DSCA-Net and other networks on lung segmentation.

Methods	Dice	Acc	Sens	Spec	Params
U-Net [10]	0.9675	0.9768	0.9441	0.9869	7.76 M
CE-Net [4]	-	0.9900	0.9800	-	-
ResU-Net (t = 2) [15]	0.9690	0.9849	0.9555	0.9945	-
RU-Net (t = 2) [15]	0.9638	0.9836	0.9734	0.9866	4.2 M
R2U-Net (t = 3) [15]	0.9826	0.9918	0.9826	0.9944	4.2 M
DSCA-Net	0.9828	0.9920	0.9836	0.9895	2.2 M

Table 6. Quantitative evaluation of ablation study on lung segmentation dataset.

Methods	Dice	IoU	ASSD
Lightweight U-Net (Backbone)	0.6160	0.4562	7.7274
Backbone + DC	0.8597	0.7661	1.0903
Backbone + DC + PA	0.9160	0.8471	1.3105
Backbone + DC + PA + CA	0.9813	0.9631	0.1192
DSCA-Net	0.9828	0.9662	0.0882

To quantitatively analyze the effectiveness, we assessed DSCA-Net with four methods: U-Net [10],

CE-Net [4], RU-Net [15], and R2U-Net [15]. Table 5 demonstrates that all methods achieve excellent performance in four metrics, and our network reached 0.9828 in Dice, 0.9920 in Acc, 0.9836 in Sens, and 0.9895 in Spec, better than U-Net. In spite of the slightly lower performance of DSCA-Net than R2U-Net in Spec, our model has 1/1.9 times fewer parameters than R2U-Net while three metric scores are higher than R2U-Net. Noting that t in Table 5 means recurrent convolution time-step.

Table 6 shows the segmentation results of ablation study on lung segmentation dataset. By adding designed modules in sequence, each of the proposed modules improved segmentation performance of DSCA-Net. The backbone + DC + PA + CA exceeds U-Net 0.0138 in Dice, and DSCA-Net shows best performance in Dice, IoU and ASSD evaluation metrics.

3.6. Nuclei segmentation

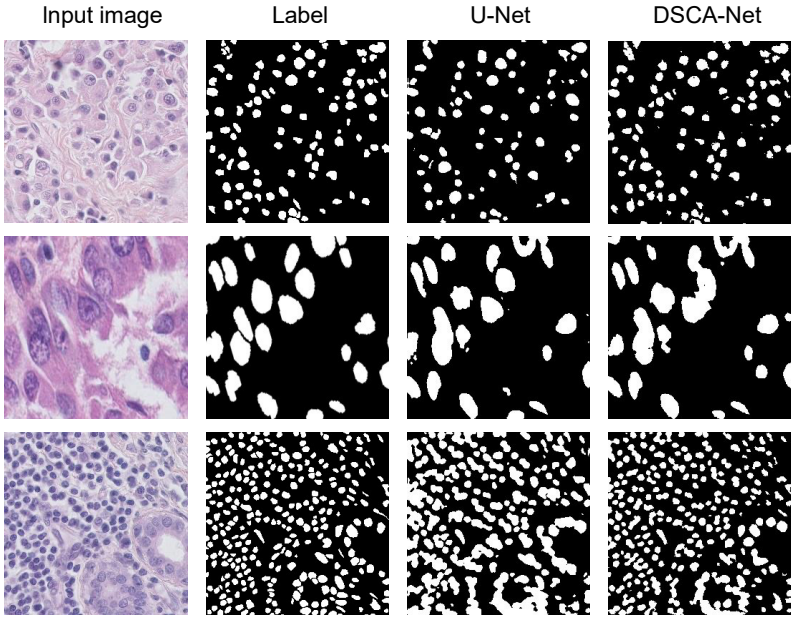

Figure 10. Visualization results of nuclei segmentation dataset.

Table 7. Comparisons of segmentation performance and number of parameters between DSCA-Net and other networks on nuclei segmentation.

Methods	Dice	Acc	Sens	Params
U-Net [10]	0.8087	0.9344	0.7915	7.76 M
DeconvNet [18]	0.8151	0.9541	0.7731	512.9 M
Ensemble [17]	0.8083	0.9441	0.9000	-
Kang et al. [19]	0.8343	-	0.8330	-
DeepLabV3+ [43]	0.8014	0.9549	-	54.7 M
Up-Net-N4 [16]	0.8369	0.9604	-	7.4 M
DSCA-Net	0.8995	0.9583	0.8544	2.2 M

The last application is nuclei segmentation of Triple-Negative Breast Cancer (TNBC) dataset. It has 50 images from 11 patients with the size of 512×512. To avoid overfitting of training process, we used data augmentation method to expand dataset with a total number of 500, including random

flipping, random cropping, and random rotation with the angle in $(-\frac{\pi}{6}, \frac{\pi}{6})$. The probability of process triggering in data augmentation methods is 0.5. As usual, we adopted the same split ratio of 7:1:2, with 350, 50, 100 for training, validation, and testing.

Figure 10 illustrates some comparative cases of prediction results between our designed network and U-Net on TNBC dataset. It can be viewed that DSCA-Net performs better than U-Net. However, incorrect segmentation results are also obtained in some segmenting areas, as shown in second line. The obscure color transitional areas and overlaid nuclei increases the difficulty to be segmented. For the relatively easy segmentation target, our network performs better.

Table 8. Quantitative evaluation of ablation study on nuclei segmentation dataset.

Methods	Dice	IoU	ASSD
Lightweight U-Net (Backbone)	0.6261	0.4562	7.7274
Backbone + DC	0.7587	0.6471	1.7348
Backbone + DC + PA	0.7691	0.6680	0.8761
Backbone + DC + PA + CA	0.8337	0.8025	0.6065
DSCA-Net	0.8995	0.8231	0.5597

Additionally, we compared DACA-Net with other networks: U-Net [10], DeconvNet [17], Ensemble [17], Kang et al. [19], DeepLabV3+ [43], and Up-Net-N4 [16]. The comparison results are shown in Table 7. Although the Sens is 0.0456 lower than Ensemble in Table 7, a combination of attention mechanism and data augmentation allows our DSCA-Net to score higher than state-of-the-art methods in Dice and Acc. Our model has 1/233.13 and 1/3.36 times fewer parameters than DeconvNet and Up-Net-N4, separately.

According to the quantitative evaluation results, Table 8 demonstrates the effectiveness of our proposed modules. After adding the MEA module, our proposed network performs better, which indicates that segmented edge is closer to the label with less error.

4. Discussion

To lighten the network parameters and maintain performance, we take fully advantages of U-Net and integrate designed modules in DSCA-Net for 2D medical image segmentation. First, DC module replaces stacked convolutional layers of U-Net for feature extraction and restoration. Second, PA module is designed to recover down-sampling feature loss. Third, CA module substitutes the simple concatenation operation in U-Net to extract richer context information. In addition, MEA module is proposed to realize segmenting target edges from multi-scale encoder information for final prediction. Evaluation metrics with other state-of-the-art networks showed the performance of DSCA-Net is better.

Multi-group experimental visualized results are shown in Figures 7–10. It can be summarized that our model is more robust than U-Net. For the blurred edge details and occlusions in electron microscope images, our network can also distinguish the segmented target correctly. For the most challenging task like TNBC, the similarity of adherent nuclei and unobvious changed color with great morphological changes increases the difficulty of segmentation. Our proposed network has achieved better results compared with other networks. However, it still needs further development.

5. Conclusions

The target of this study is to lighten parameters of the network while maintaining good performance. We design a lightweight depthwise separable convolutional neural network with an attention mechanism named DSCA-Net for accurate medical image segmentation. Our proposed network extracts richer feature information and reduces feature loss in segmentation processing compared with U-Net. We assessed our network on four datasets and collected segmentation results against state-of-the-art networks under various metrics. The visualization and quantitative results show that our network has better segmenting ability. We intend to utilize DSCA-Net to segment 3D images in the future.

Acknowledgments

This study was supported by Shanghai Jiao Tong University Medical-industrial Cross-key Project under Grant ZH2018ZDA26, the Jiangsu Provincial Key Research and Development Fund Project under Grant BE2017601.

Conflict of interest

The authors have no conflicts of interest to declare.

References

1. J. C. Caicedo, J. Roth, A. Goodman, T. Becker, K. W. Karhohs, M. Broisin, et al., Evaluation of deep learning strategies for nucleus segmentation in fluorescence images, *Cytometry, Part A, J. Quant. Cell Sci.*, **95** (2019), 952–965. https://doi.org/10.1002/cyto.a.23863
2. Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, X. Yang, A review of deep learning based methods for medical image multi-organ segmentation, *Physica Med.*, **85** (2021), 107–122. https://doi.org/10.1016/j.ejmp.2021.05.003
3. R. Merjulah, J. Chandra, Segmentation technique for medical image processing: a survey, in *2017 International Conference on Inventive Computing and Informatics (ICICI)*, (2017), 1055–1061. https://doi.org/10.1109/ICICI.2017.8365301
4. Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, et al., CE-Net: context encoder network for 2D medical image segmentation, *IEEE Trans. Med. Imaging*, **38** (2019), 2281–2292. https://doi.org/10.1109/TMI.2019.2903562
5. N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, 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 *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (2018), 168–172. https://doi.org/10.1109/ISBI.2018.8363547
6. H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, M. J. Deen, Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives, *Neurocomputing*, **444** (2021), 92–110. https://doi.org/10.1016/j.neucom.2020.04.157

7. N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, X. Ding, Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation, *Med. Image Anal.*, **63** (2020), 101693. https://doi.org/10.1016/j.media.2020.101693
8. V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615
9. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015), 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965
10. O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
11. R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, et al., CA-Net: comprehensive attention convolutional neural networks for explainable medical image segmentation, *IEEE Trans. Med. Imaging*, **40** (2021), 699–711. https://doi.org/10.1109/TMI.2020.3035253
12. N. Ibtehaz, M. S. Rahman, MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation, *Neural Networks*, **121** (2020), 74–87. https://doi.org/10.1016/j.neunet.2019.08.025
13. C. Y. Chang, P. C. Chung, Y. C. Hong, C. H. Tseng, A neural network for thyroid segmentation and volume estimation in CT images, *IEEE Comput. Intell. Mag.*, **6** (2011), 43–55. https://doi.org/10.1109/MCI.2011.942756
14. S. Nandamuri, D. China, P. Mitra, D. Sheet, Sumnet: fully convolutional model for fast segmentation of anatomical structures in ultrasound volumes, in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, (2019), 1729–1732. https://doi.org/10.1109/ISBI.2019.8759210
15. M. Z. Alom, C. Yakopcic, T. M. Taha, V. K. Asari, Recurrent residual convolutional neural network based on U-Net (R2U-Net) for medical image segmentation, in *NAECON 2018 - IEEE National Aerospace and Electronics Conference*, (2018), 228–233. https://doi.org/10.1109/NAECON.2018.8556686
16. Y. Wen, L. Chen, Y. Deng, J. Ning, C. Zhou, Towards better semantic consistency of 2D medical image segmentation, *J. Visual Commun. Image Represent.*, **80** (2021), 103311. https://doi.org/10.1016/j.jvcir.2021.103311
17. P. Naylor, M. Lae, F. Reyal, T. Walter, Nuclei segmentation in histopathology images using deep neural networks, in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, (2017), 933–936. https://doi.org/10.1109/ISBI.2017.7950669
18. H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in *2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1520–1528. https://doi.org/10.1109/ICCV.2015.178
19. Q. Kang, Q. Lao, T. Fevens, Nuclei segmentation in histopathological images using two-stage learning, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, (2019), 703–711. https://doi.org/10.1007/978-3-030-32239-7_78

20. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a nested U-Net architecture for medical image segmentation, in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, **2018** (2018), 3–11. https://doi.org/10.1007/978-3-030-00889-5_1
21. L. Kaiser, A. N. Gomez, F. Chollet, Depthwise separable convolutions for neural machine translation, preprint, arXiv:1706.03059.
22. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
23. Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, ECA-Net: efficient channel attention for deep convolutional neural networks, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 11531–11539. https://doi.org/10.1109/CVPR42600.2020.01155
24. V. Mnih, N. Heess, A. Graves, K. Kavukcuoglu, Recurrent models of visual attention, in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, **2** (2014), 2204–2212. https://dl.acm.org/doi/abs/10.5555/2969033.2969073
25. D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, preprint, arXiv:1409.0473.
26. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attenti on is all you need, in *Advances in Neural Information Processing Systems*, **30** (2017), 5998–6008. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
27. O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention U-Net: learning where to look for the pancreas, preprint, arXiv:1804.03999.
28. F. Chollet, Xception: deep learning with depthwise separable convolutions, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1800–1807. https://doi.org/10.1109/CVPR.2017.195
29. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, et al., Mobilenets: Efficient convolutional neural networks for mobile vision applications, preprint, arXiv:1704.04861.
30. X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: an extremely efficient convolutional neural network for mobile devices, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018), 6848–6856. https://doi.org/10.1109/CVPR.2018.00716
31. L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *Computer Vision – ECCV 2018*, (2018), 833–851. https://doi.org/10.1007/978-3-030-01234-2_49
32. K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, et al., X-Net: brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, (2019), 247–255. https://doi.org/10.1007/978-3-030-32248-9_28
33. A. Wibowo, S. R. Purnama, P. W. Wirawan, H. Rasyidi, Lightweight encoder-decoder model for automatic skin lesion segmentation, *Inf. Med. Unlocked*, **25** (2021), 100640. https://doi.org/10.1016/j.imu.2021.100640

34. C. Meng, K. Sun, S. Guan, Q. Wang, R. Zong, L. Liu, Multiscale dense convolutional neural network for DSA cerebrovascular segmentation, *Neurocomputing*, **373** (2020), 123–134. https://doi.org/10.1016/j.neucom.2019.10.035
35. G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
36. Y. Wu, K. He, Group normalization, *Int. J. Comput. Vision*, **128** (2020), 742–755. https://doi.org/10.1007/s11263-019-01198-w
37. X. Zhang, Y. Zou, W. Shi, Dilated convolution neural network with LeakyReLU for environmental sound classification, in *2017 22nd International Conference on Digital Signal Processing (DSP)*, (2017), 1–5. https://doi.org/10.1109/ICDSP.2017.8096153
38. P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, *Sci. Data*, **5** (2018), 180161. https://doi.org/10.1038/sdata.2018.161
39. T. Wunderling, B. Golla, P. Poudel, C. Arens, M. Friebe, C. Hansen, Comparison of thyroid segmentation techniques for 3D ultrasound, *Med. Imaging 2017: Image Process.*, **10133** (2017), 1013317. https://doi.org/10.1117/12.2254234
40. D. P. Kingma, J. Ba, Adam: a method for stochastic optimization, preprint, arXiv:1412.6980.
41. G. Lin, A. Milan, C. Shen, I. Reid, RefineNet: Multi-path refinement networks for high-resolution semantic segmentation, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 5168–5177. https://doi.org/10.1109/CVPR.2017.549
42. R. Ma, S. Zhang, C. Gan, H. Zhao, EOCNet: Improving edge omni-scale convolution networks for skin lesion segmentation, in *2020 3rd International Conference on Digital Medicine and Image Processing*, (2020), 45–50. https://doi.org/10.1145/3441369.3441377
43. L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, *IEEE Trans. Pattern Anal. Mach. Intell.*, **40** (2018), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184
44. S. Chen, Y. Zou, P. X. Liu, IBA-U-Net: Attentive BConvLSTM U-Net with redesigned inception for medical image segmentation, *Comput. Biol. Med.*, **135** (2021), 104551. https://doi.org/10.1016/j.compbiomed.2021.104551

AIMS Press