SUN, M., LI, P., REN, J. and WANG, Z. 2023. Attention mechanism enhanced multi-layer edge perception network for deep semantic medical segmentation. *Cognitive computation* [online], 15(1), pages 348-358. Available from: https://doi.org/10.1007/s12559-022-10094-4

Attention mechanism enhanced multi-layer edge perception network for deep semantic medical segmentation.

SUN, M., LI, P., REN, J. and WANG, Z.

2023

Attention Mechanism Enhanced Multi-layer Edge Perception Network for Deep Semantic Medical Segmentation

Meijun Sun[1,2] · Pengfei Li[1,2] · Jinchang Ren[3] · Zheng Wang[1,2]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China

[2] Tianjin Key Lab of Machine Learning, Tianjin, China

[3] National Subsea Centre School, Robert Gordon University,

Corresponding author: Zheng Wang, email: wzheng@tju.edu.cnAberdeen, UK

Received: 16 June 2022 / Accepted: 4 December 2022

Abstract

Existing deep learning–based medical image segmentation methods have achieved gratifying progress, but they still suffer from the coarse boundaries with similar pixels of target. Because the boundary of medical images becomes blurred and the gradient is inconsistent and not apparent, high-resolution images are needed for more accurate segmentation. To tackle these problems, we propose an efficient multi-layer edge perception U-shaped structure for medical image segmentation. In this paper, we present a multi-layer edge perception network for describing more precise edges of medical targets. The U-structure architecture of our network embeds a multi-layer edge perception module, which has the following advantages: (1) connect-ing different scales and channels to help the network better learn the feature of the medical image via the combination of a pyramid structure and several edge perception modules; (2) a new downsampling block is designed to improve the network's sensibility to the target boundary. We demonstrate the effectiveness of the proposed model on the DRIVE datasets, and achieve a Dice gain of 0.841 over other models. In this paper, we propose an efficient multi-layer edge perception U-shaped structure for medical image segmentation. A large number of experiments show that the performance of our proposed multi-layer edge perception U-shaped network is significantly better than the traditional segmented network structure.

Keywords Deep medical segmentation, U-Structure network, Attention mechanism, Semantic segmentation

Introduction

As one of the basic themes of computer vision, semantic segmentation aims to assign semantic labels to each pixel in the image [1], which has been successfully applied in many fields, e.g., augmented reality and autonomous driving [2, 3], and human-machine interaction. With a full convolutional network (FCN) [4], the performance of deep convolutional network in the segmentation task has been significantly improved. For medical image segmentation, U-Net [5] has greatly reduced the size of the dataset for training the neural network yet still further improves the segmentation performance.

As the internal structure of the human body is relatively fixed, the distribution of the targets for segmentation is similar, where the semantics are simple and quite clear. However, this may be affected by the information of the images, which may cause the blurred view of the targets. At the same time, the boundary of medical images becomes blurred and the gradient is inconsistent and not apparent; thus, high-resolution images are needed for more accurate segmentation. Therefore, the encoder-decoder structure [6–8] used in U-Net provides a good solution for the subsequent medical image segmentation.

For medical image segmentation, the full convolutional network (FCN) [4] and U-Net [5] are the two commonly used architectures. Despite their good representational power, they are still incapable of extracting enough information of target edge. To tackle this issue, researchers have made a lot of research from different aspects, such as using different scale feature maps [9–12] for enhanced feature extraction, the addition of the attention mechanism [13–15] for improved feature representation, and optimizing the loss function [16–19] can be used not only for speed but also bet-ter performance to accelerate the convergence of the network while training.

In this paper, we also use the encoder-decoder structure for semantic segmentation, which can extract features through the cascade of convolution modules. During the convolution, a low-level feature map contains more details for boundary segmentation, while a high-level feature map has more abstract features for positioning [20]. Here comes a challenging problem as how to combine the information from the low- and high-level feature maps for improved edge perception. To tackle this problem, we propose an edge perception module which can efficiently extract both the high-level and low-level features and fuse them. At the same time, some information may be lost in the processes of upsampling and downsampling, which requires us to adopt some measures to reduce such information loss [16]. In order to further improve the segmentation accuracy and reduce the information loss, we also redesign the downsampling block.

In summary, the main contributions of our paper are threefold:

- We propose a multi-layer edge perception module for embedding in a U-structure by combining an image pyramid with the attention mechanism for improving feature extraction and edge preservation. In addition, the proposed module can be extended to other network models for enhancing their ability of feature representation.
- We redesigned the downsampling convolution block, inspired by the idea of the residual network [21], which enables the encoder to more effectively capture fine-grained details of the foreground objects and reconstructed in the decoder.
- We have implemented a useful extension of the standard Net, by combining the multi-layer edge perception and downsampling convolution block. On the DRIVE datasets, we achieve a Dice gain of 0.841 over other models. It provides a new research idea for medical image segmentation.

Related Works

Encoder-Decoder The encoder-decoder structure is first applied in the field of natural language processing [22, 23]. Since it can effectively encode and convert the information, it is later widely used in the field of computer vision. This structure consists of two parts. In the encoder, the spatial dimension of the feature map is continuously reduced, and the larger range of information is more easily obtained at the deeper level. In the decoder, the details and spatial dimensions of the object are restored. Such as in [4, 24], the deconvolution [25] method is used to upsample the low-resolution features; in U-NET [5], the jumping connection method is used to fuse the features extracted by the encoder with the upsampled information of the decoder. The relevant works [26–29] have demonstrated the validity of the encoder-decoder–based structural in several semantic segmentation benchmarks. In addition, the encoder-decoder structure is also applied to salient target detection [30, 31].

Attention Mechanism The mechanism of attention originated from the research on human vision, which was initially applied in the field of machine translation, and then gradually became an important part of neural networks, and its application fields were expanded to natural language processing, computer vision, and other aspects. Attention mechanism is a means to quickly screen out high-value information from a large amount of information with limited attention resources. It can be understood that it redistributes the computing resources which were originally distributed equally according to the importance of attention objects, and concentrates the resources to the part with the most abundant information. In the related research of computer vision, the core idea of attention mechanism is to find correlations based on the original data, so as to highlight some important features. The spatial domain attention proposed by Jaderberg et al. [32] can transform the spatial information in the original image to another space while retaining the key information. The channel attention proposed by Hu et al. [14] can allocate resources among various convolution channels. Wang et al. [33] proposed combining attention mechanism with the mask that put forward the residual attention to learn not only the mask after the characteristics of the tensor as input of the next layer. At the same time, it will mask the characteristics of the tensor as input of the next layer; this way it can get more abundant characteristics, and thus better able to pay attention to the key characteristics.

Proposed Method

Architecture Overview

As shown in Fig. 1, the proposed multi-layer edge perception network is a U-shaped structure, which is widely used in the field of semantic segmentation [13, 30]. It consists of encoder and decoder. In the encoder, there are many repeated multi-path downsampling block, and each block is followed by a ReLU activation function and a maximum pooling layer with a step size of 2. In the decoder, the inter-polation algorithm is used for the upsampling operation. The result is connected with the corresponding feature map to improve the network generalization ability.

We introduce two innovative modules into the network architecture :(1) The multi-layer edge perception module combines the pyramid idea with the attention mechanism, which can make the network more focused on sensitive areas, thus improving the ability of the network to capture feature information and the capacity of segmenting the edge of objects. (2) We design a multi-path downsampling block to guide the extraction and fusion of multiscale information, which can make the multi-level segmentation information obtained by the network, so as to make up for the loss of information caused in the process of multiple downsampling.

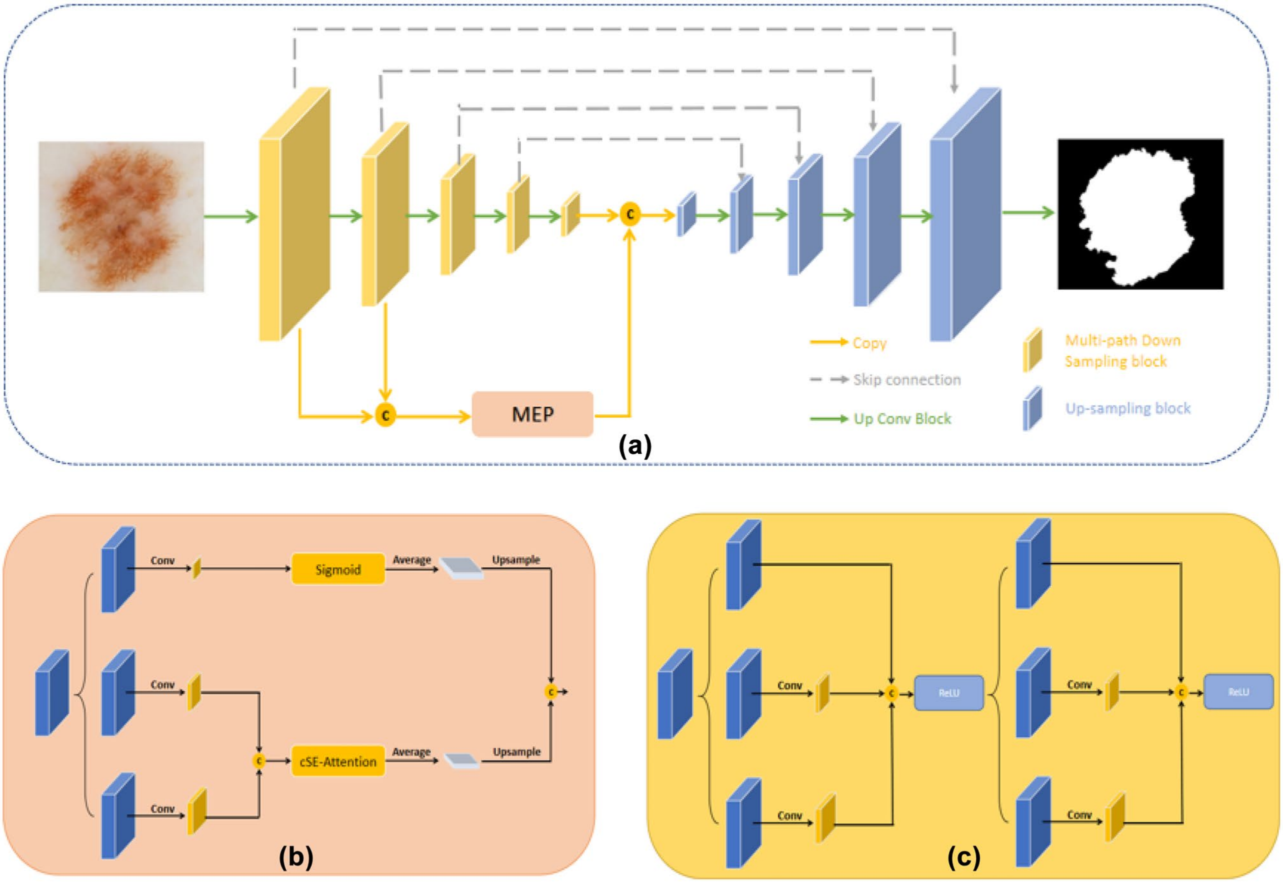

Fig. 1 Illustration of the proposed multi-layer edge perception module U-shaped network. There are three modules,respectively, the MEP network (**a**), the multi-layer edge perception module (**b**), and the multi-path downsampling module (**c**). The input image goes through the multi-path downsampling module first in which the downsampled factor is 2 at each scale. The result from module c concated as the input of multi-layer edge perception module (b). After that, we concate their result and the final segmentation result is obtained after several interpolation upsampling

To be more specific, we first deploy multi-layer edge perception module and multi-path downsampling block in the encoder to extract a series of depth feature maps. Then, the upsampling operation is carried out by interpolation algorithm in the decoder to restore the image resolution. Finally, we convolve the output result of the decoder to get the segmentation result.

Multi-layer Edge Perception Module

The key to make high-quality semantic segmentation is that network can capture enough detail information. At this point, the researchers have done a lot of work, such as Hu et al. [14] who proposed a cross-channel attention mechanism, modeling the features based on the dependence of each channel in order to improve the network express ability. Jaderberg et al. [32] proposed the spatial attention mechanism that can select the region of interest, and find the areas that need to be paid more attention in the features. The pyramid structure proposed by He et al. [34] extracts information from feature maps of various dimensions, which enables the network to capture more abundant features and improve the network's ability to understand the semantic information.

Based on the above three advantages of design, we put forward a module named multi-layer edge perception module, which has three different sizes of convolution layers to extract multiple figures of features of the medical image information, and at the same time we introduce spatial atten-tion mechanism and cross-channel attention mechanism to enhance the sensitivity of the network to object boundaries. We fuse the multi-level feature extraction which establishes dependencies between channels. And then we homogenize the result to selectively reinforce features containing useful information and suppress unwanted features. Finally, we upsample the feature and fuse them as our result which is the input of the decoder.

To be more specific, we adopt the idea of pyramid structure model in the multi-layer edge perception module framework. Input features are extracted in different dimensions through a 3*3 convolution layer and a 5*5 convolution layer respectively. Then, the extracted feature information is fused and put into the cross-channel attention module to establish the channel relationship dependency model. In the meantime, the input feature map is also sent into a 1*1 con-volution layer and sigmoid function. The two outputs are concated as the result after being re-coded. This module not only enhances the network feature representation, but also strengthens the network's description of edge details. The multi-layer edge perception module framework is shown in Fig. 2.

CSE Module

In order to further improve the capability of the network and reduce the parameter, we propose a new channel attention model named CSE module, which is more effective than the original SE module. The SE module is a representative channel attention mechanism module in the neural network. It can enhance the representation of features by modeling the dependencies between feature mapping channels. The SE module first compresses the spatial dependencies through spatial global average pooling to learn channel-specific feature representation, then using two full connection layers to compress and increase the dimensions of channels, and finally through a sigmoid function to highlight the useful channels. Assume that given the feature map $X_i \in R^{C*W*H}$, the channel attention feature map $A_{ch}(X_i) \in R^{C*1*1}$ is calculated by the following formula:

$$A_{ch}(X_i) = \sigma(W_C(\delta(W_{\frac{C}{r}}(F_{gap}(X_i)))))$$

where $F_{gap}(X) = \frac{1}{WH} \sum_{i,j=1}^{W,H} X_{i,j}$ is to calculate the global average pooling of each channel, $W_{\frac{C}{r}} \in R^{C*1*1}$ and $W_c \in R^{C*1*1}$ represent the parameters of the full connection layer in the process of dimension reduce and dimension increase, respectively, σ represents the sigmoid activation function, and δ represents the ReLU function.

However, the SE module also has its problem. The first is the loss of information due to channel compression in the process of dimensionality reduction. In the SE module, the input feature map will first go through a fully connected layer for dimensionality reduction, and then the original dimension will be restored by another fully connected layer. The reduction of channel dimension will cause partial loss of information [35]. In addition, due to the fully connected layer, with the deepening of network depth, the number of channels becomes higher and the number of parameters becomes larger, which will cause a large computational burden, resulting in reducing the convergence rate of network and making training difficult.

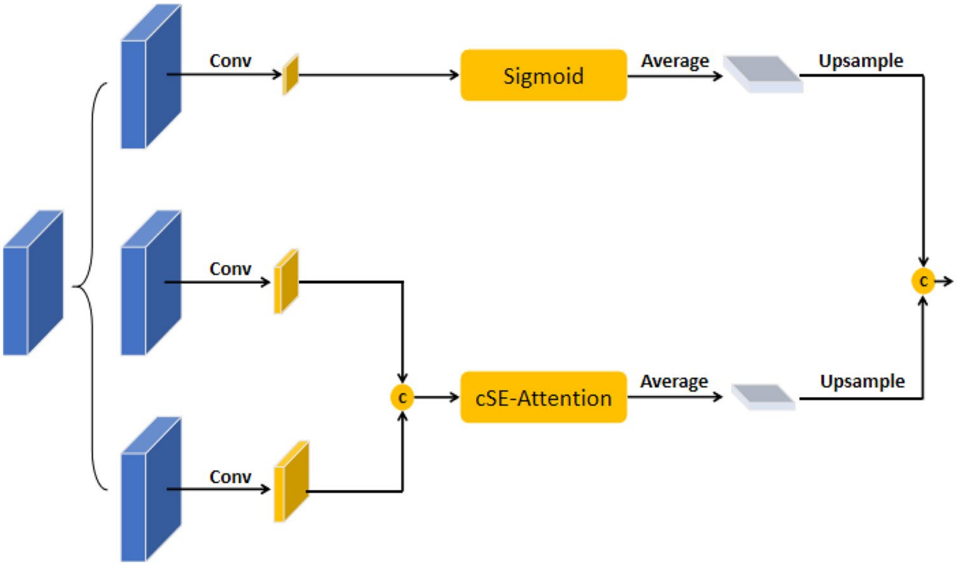

Fig. 2 Multi-layer edge perception module. The input goes through three convolution layers with kernel sizes of 1*1, 3*3, and 5*5, respectively. The output of 1*1 convolution layer is fed into the sigmoid function. The outputs of 3*3 and 5*5 convolution layer are fused and then sent into the channel attention module. The results were homogenized and then upsampled. We fuse the two results as the output

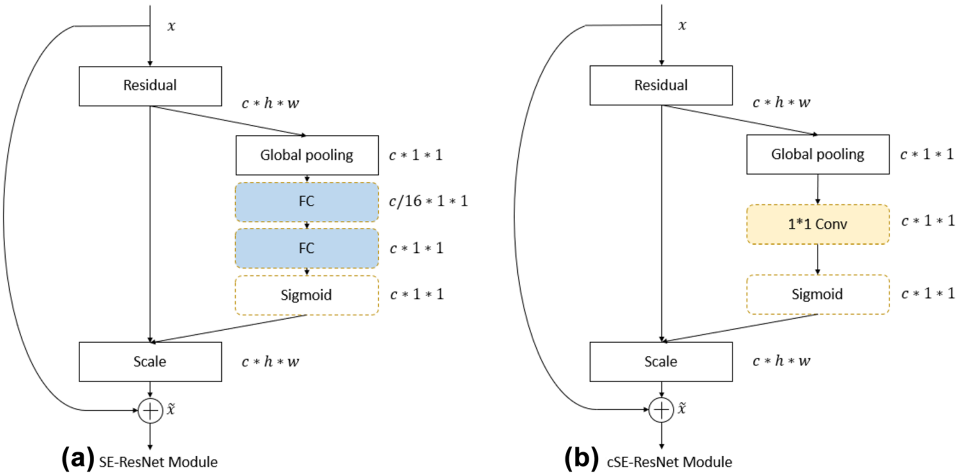

Fig. 3 The structure of SE module and CSE module: **a** traditional SE module, **b** CSE module. In the traditional SE module, two layers of a full connection layer are used for dimension reduction and dimension increase, respectively, causing the loss of information. In the CSE module, use a layer of convolution instead. FC means fully connected

Therefore, we propose to optimize the SE module by using the convolution layer, replacing the original two-layer fully connected network with one layer of convolution, which also can effectively establish the relationship dependence between channels while reducing the loss of information. As shown in Fig. 3, suppose that given the feature map $X_i \in R^{C*W*H}$, the channel attention feature map $A_{ch}(X_i) \in R^{C*1*1}$ is calculated by the following formula:

$$A_{ch}(X_i) = \sigma(W_C(F_{gap}(X_i)))$$

where, $F_{gap}(X) = \frac{1}{WH}\sum_{i,j=1}^{W,H} X_{i,j}$ calculates the global average pooling of each channel, $W_c \in R^{C*1*1}$ represents the parameters of the channel-dependent convolutional layer, and σ represents the sigmoid activation function.

Multi-path Downsampling Block

Inspired by the deep residual network [21], pooled pyramid model [34], and DenseNet [36], we proposed a new downsampling block which fully utilizes the advantages of the above three models. The residual network simplifies the learning process and enhances the gradient propagation during training the network by way of skip connection. Meanwhile, studies show [37] that the use of residual structure breaks network asymmetry, reduces neural network degradation, and enhances network generalization ability. The pooling pyramid model can use different sizes of convolution kernels to obtain different scales of feature maps so as to obtain different dimensions of information and enhance the representational ability of the network. DenseNet makes use of intensive skip connections to full use of features, which makes up for the information loss caused in the downsampling process and makes the network have a smaller generalization error bound [38].

As shown in Fig. 4, the proposed multi-path downsam-pling block carries out 3*3 convolution and 1*1 convolution for the input, respectively, and fuses the convolution result with the input jump connection as the output. Assuming that the input of layer l is x_l, the output after 3*3 convolution is $F_{3l}(x_l)$, and the output after 1*1 convolution is $F_{1l}(x_l)$, then the output of layer l is expressed as follows:

$$O_l = F_{3l}(x_l) + F_{1l}(x_l) + x_l$$

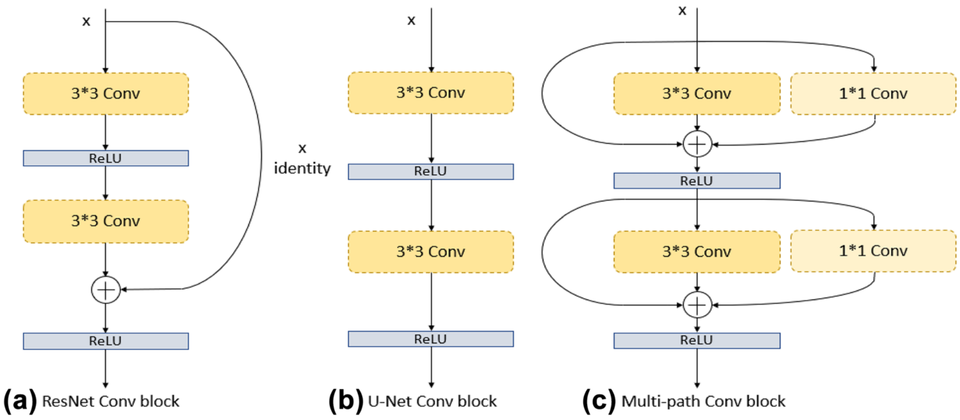

Fig. 4 Different ways of downsampling block: **a** ResNet conv block, **b** U-Net conv block, **c** multi-path Conv block

The advantages of using the proposed downsampling block are as follows: (a) it can obtain feature maps of different dimensions through the different scales of convolution layers, so that the network can capture more global feature information through different sizes of receptive fields; (b) it can accelerate the convergence of the network and improve the generalization ability of the network through dense jump connections.

Experiment

Experiment Details

Dataset For the experiments, the ISIC Archive dataset is used. This dataset is publicly available and commonly used to medical image segmentation frameworks. This ISIC Archive contains over 23K images of skin lesions, labeled "benign" or "malignant." In the dataset, some images have multiple segmentation offered, made with different skill levels. We randomly pick up 3K images as the training set and 1K images as the testing set.

Data Augmentation For data augmentation, we adopt random mirror and random resize between 0.5 and 2 for all datasets, and additionally add random rotation between −10 and 10°, and random Gaussian blur for benchmark. This comprehensive data augmentation scheme makes the network resist overfitting [9].

Training Details All the models were trained on a single NVIDIA 1080Ti with 11GB memory. During the training process, parameters were initialized randomly, and RMSProp algorithm was used as the optimizer, and the initial learning rate was 0.0001. Unless specified, the input image is resized to 800*800 pixels, using a weight decay factor of 1e−8 and a momentum drop parameter of 0.9, respectively.

Evaluation Metrics We evaluate our method by 4 widely used metrics: Dice coefficient (Dice), pixel accuracy (PA), sensitivity (Sen), and volumetric overlap error (VOE). Con-sidering tumor segmentation region (positive) and background (negative), we compute the terms true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

1. Dice coefficient was used as the evaluation index, which is usually used to calculate the similarity of two samples. The calculation formula of Dice coefficient is as follows:

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

2. Pixel accuracy is the ratio of all correctly classified pixels to the total number of pixels. The calculation formula is as follows:

$$PA = \frac{TP + TN}{TP + TN + FP + FN}$$

3. Sensitivity is the ratio of the number of samples correctly predicted as positive to the total number of true positive samples; it is computed as follows:

$$Sensitivity = \frac{\mathrm{TP}}{TP + FN}$$

4. VOE means volumetric overlap error; it is computed asfollows:

$$VOE = 1 - \frac{TP}{FN + FP - TP}$$

Ablation Study

Multi-layer Edge Perception Module Table 1 shows the results between U-Net and U-Net with the multi-layer edge perception module. In the edge perception module, we introduce channel and spatial attention mechanism to extract medical image information from different dimensions. By recoding the features, we can selectively strengthen the features containing useful information and suppress the useless features.

To explore more possibilities, we also conduct experiments on different interpolation functions and sub-pixel convolution methods. For different interpolation algorithms, there is no significant difference between them on this dataset. At the same time, the effect of sub-pixel convolution is also not improved. We think that may be caused by the image pixel quality. From Table 1, it can be seen that compared with the traditional U-Net model, our method makes the Dicecoefficient increase by 0.03. At the same time, for other methods, our proposed module also plays a good effect.

Table 1 Ablation experiment results for validation of multi-layer edge perception modules

Method	PA	Sen	VOE	Dice
U-Net	0.927	0.634	0.371	0.762
U-Net+MEP(SE)+SPC	0.931	0.669	0.334	0.787
U-Net+MEP(SE)+NNI	0.946	0.685	0.308	0.793
U-Net+MEP(SE)+BI	0.944	0.689	0.311	0.795

MEP multi-layer edge perception module, *SE* cross-channel attention module, *BI* bilinear interpolation, *NNI* nearest neighbor interpolation, *SPC* sub-pixel convolutional

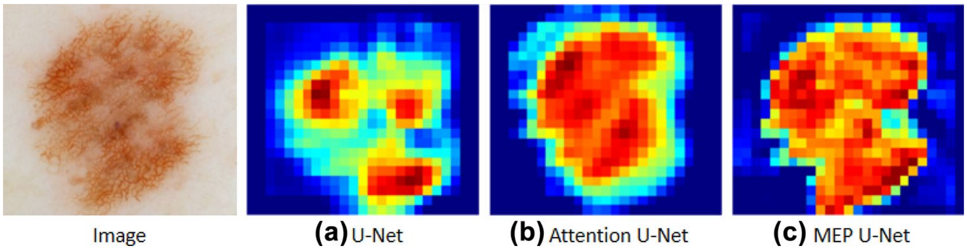

Fig. 5 Comparison of heat maps of various networks. The image is the actual image: **a** is the heat map of the last downsampling block, **b** is the heat map of the last attention module, **c** is the heat map of the multi-layer edge perception module

In order to better show the effectiveness of the multi-layer edge perception module, we demonstrate the heat map of different U-structure networks. The picture comes from the DRIVE datasets. It can be seen that our method has more focus on the boundary of the object compared to the other U-structure method in Fig. 5. Especially to the network with the single attention module, the proposed method combined with the idea of multiple attention mechanisms can improve the detail presentation ability in the picture. That is because through the MEP module, our network can locate more information. When the network is propagating forward, the invalid information has been suppressed. Thus, the network can segment the object edge details better and more accurate segmentation results can be obtained.

CSE Module Table 2 shows the result between MA-UNet with SE module, ECA module, and CSE module. Compared with the traditional SE module, we use a single convolution layer to replace the two fully connected layers in the SE module. The idea is somewhat similar to ECANet, so we also conducted experimental comparisons with ECA module. From Table 2, it can be seen that our optimization method can improve the Dice coefficient by 0.013 compared with the traditional SE module. Compared with the ECA module, our method also achieved relatively good results. We think that this is due to the fact that for datasets with small pixel quality, it may be redundant to establish the connection between channels by multi-layer convolution.

In order to further explore the effect of different convolution kernel sizes on performance, we also set the CSE convolution layer sizes of 1*1, 3*3, and 5*5 for experiments. As can be seen in Table 3, as the size of convolution kernel increases, the segmentation effect decreases. This is because the size of features in the CSE module is small, and it is easy to lose a lot of information if a large convolution kernel is used at this time, so the kernel size of 1*1 is the best choice.

Table 2 Ablation experiment for validation of the CSE module compared with ECA module and SE module. The CSE module uses a layer of convolution to replace the two fully connected layers in the SE module

Method	PA	Sen	VOE	Dice
U-Net	0.927	0.634	0.371	0.762
U-Net+MEP(ECA)	0.941	0.673	0.322	0.781
U-Net+MEP(SE)	0.944	0.689	0.311	0.795
U-Net+MEP(CSE size=1)	0.951	0.705	0.291	0.808

Table 3 Ablation experiment for CSE module with different sizes of kernel

Method	PA	Sen	VOE	Dice
U-Net	0.927	0.634	0.371	0.762
U-Net+MEP(CSE size=5)	0.931	0.611	0.388	0.756
U-Net+MEP(CSE size=3)	0.934	0.655	0.356	0.771
U-Net+MEP(CSE size=1)	0.951	0.705	0.291	0.808

Multi-path Downsampling Block Table 4 shows the result between the different methods with the multi-path downsampling block. Compared with the traditional downsampling block, we use three different sizes of convolution layers to capture feature information. Meanwhile, we add intensive skip connection to improve the convergence and generalization ability of the network. According to Table 4, it can be seen that compared with the traditional downsampling block, our multi-path downsampling block makes the Dice coefficient increase by 0.043 with U-Net. It also works in ResNet. That is because the multi-path downsampling block can obtain feature maps of different dimensions, so that the network can capture more feature information to make up for the loss of information in the process of downsampling.

Table 4 Ablation experiment for validation of multi-path downsampling block

Method	PA	Sen	VOE	Dice
ResNet	0.930	0.592	0.410	0.711
ResNet+MDS	0.933	0.622	0.398	0.732
U-Net	0.927	0.634	0.371	0.762
U-Net+MDS	0.959	0.718	0.282	0.805
U-Net+MEP+CSE+MDS	0.963	0.754	0.249	0.841

MDS multipath downsampling

Table 5 Comparison of our method and the seven methods. The listed segmentation frameworks are evaluated on the same public dataset using the same number of training and testing images

Method	Parameters	PA	Sen	VOE	Dice
R2U-Net [39]	156.5M	0.932	0.578	0.422	0.706
Deeplabv3+ [12]	161.7M	0.935	0.697	0.387	0.710
Attention R2U-Net [39]	157.9M	0.921	0.617	0.401	0.733
U-Net [5]	69.1M	0.927	0.634	0.371	0.762
U-Net++ [7]	36.7M	0.936	0.678	0.326	0.798
Attention U-Net [13]	92.7M	0.942	0.701	0.299	0.802
U-Net+MDS	73.4M	0.959	0.718	0.282	0.805
FCN [4]	60.5M	0.971	0.859	0.261	0.818
MEP-UNet (ours)	111.1M	0.963	0.754	0.249	0.841

We demonstrated the effectiveness of our module through a series of ablation experiments. At the same time, in the experimental process, we can also find that each module can still play a good mutual enhancement effect.

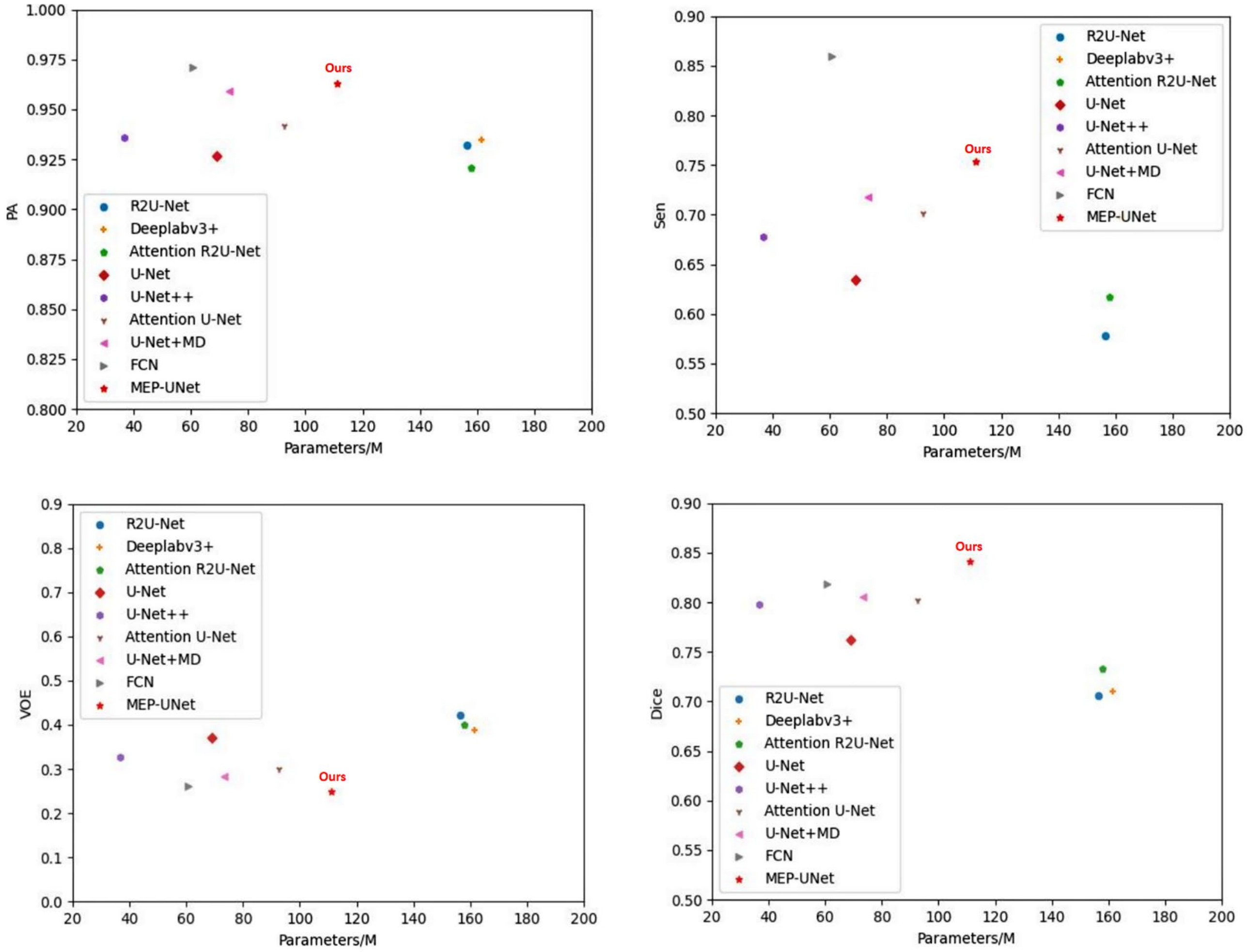

Fig. 6 Scatter plots of four indicators with different models

Comparison with Other Medical Segmentation Method

We compare our model with a series of U-structure methods including a one-base model, U-Net, and four U-structure models: U-Net++, R2U-Net, Attention U-Net, Attention R2U-Net. At the same time, we also compare with networks based on ResNet like Deeplabv3+ and FCN.

Quantitative Comparison

Results from the U-structure segmentation method are summarized in Tables 4 and 5 for comparison purposes. Because these methods have not been trained on the DRIVE datasets before, we retrained them on the same training datasets with the same auxiliary function; this comparison gives an insight on how our proposed method compares to the relevant work.

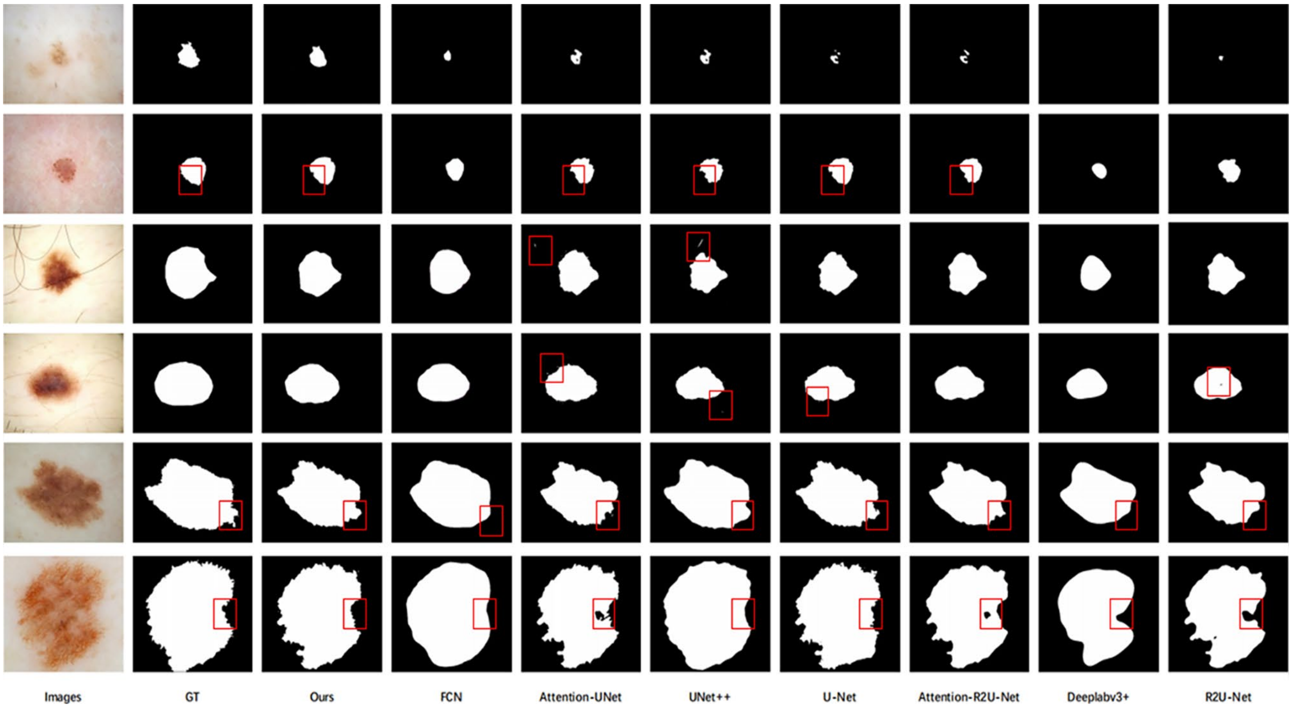

Fig. 7 Qualitative comparison of the proposed method with five other U-structure methods. From the right to left are Images, GT, Ours, FCN, Attention-UNet, UNet++, U-Net, Attention-R2U-Net, Deeplabv3+, R2U-Net

To more clearly compare the advantages of each model, we make the scatter plots of four indicators with different models. From Table 4 and Fig. 6, it can be seen that the single attention model does not maximize the power of the attention mechanism. In the process of downsampling, simply deepening the depth and adding the parameters of the network do not work for the result. Lastly, our method got better performance on ISIC Archive, which achieved 0.841 Dice.

Qualitative Comparison

To give an intuitive understanding of the promising performance of our models, we illustrate the sample results of our models and several other methods in Fig. 7. It can be seen that our MEP-UNet is able to divide the target edge more clearly.

The rows 1 through 4 of Fig. 7 show the results of different shapes of skin melanoma. As we can observe, our MEP-UNet is able to produce accurate results on the large and small skin melanoma. We also can see from the third and fourth rows of Fig. 7 that our method can achieve high-quality segmentation of the object disturbed with hair. The fifth and sixth rows show the results of target with complex boundaries. Compared to the single attention mechanism, our method can more focus on the boundary of the object. In summary, our model can handle various skin melanoma and produce high-accuracy segmentation results.

Conclusions

In this paper, we propose an efficient multi-layer edge perception U-shaped structure for medical image segmentation. We use multi-path downsampling block to extract rich information and fuse them to put into a multi-layer edge perception module which combined the advantages of pooling pyramid and attention mechanism. It can more efficiently use the global information to selectively strengthen the features containing useful information and suppress the useless features. At the same time, it can make the second refinement of the segmentation object edge. A large number of experiments show that the performance of our proposed multi-layer edge perception U-shaped network is significantly better than that of the traditional segmented network structure.

Funding This work was supported by the National Natural Science Foundation of China (CN) (Nos. 61876125, and 62076180).

Data Availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest The authors declare no competing interests.

References

1. Minaee S, Boykov YY, Porikli F, Plaza AJ, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. IEEE Trans Pattern Anal Mach Intell 2021;1–1. https://doi.org/10.1109/TPAMI.2021.3059968.
2. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
3. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? the kitti vision benchmark suite, in. IEEE Conference on Computer Vision and Pattern Recognition. 2012;2012:3354–61. https://doi.org/10.1109/CVPR.2012.6248074.
4. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015.
5. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41.
6. del Mar Vila M, Remeseiro B, Grau M, Elosua R, Betriu À, Fernandez-Giraldez E, Igual L. Semantic segmentation with densenets for carotid artery ultrasound plaque segmentation and CIMT estimation. Artif Intell Med. 2020;103:101784. https://doi.org/10.1016/j.artmed.2019.101784. https://www.sciencedirect.com/science/article/pii/S093336571830770X
7. Zhou Z. M. M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, Tavares JMR, Bradley A, Papa JP, Belagiannis V, Nascimento JC, Lu Z, Conjeti S, Moradi M, Greenspan H, Madabhushi A, editors. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Cham: Springer International Publishing; 2018. p. 3–11.
8. Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell. 2017;39(12):2481–95. https://doi.org/10.1109/TPAMI.2016.2644615.
9. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
10. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans Pattern Anal Mach Intell. 2018;40(4):834–48. https://doi.org/10.1109/TPAMI.2017.2699184.
11. Chen L-C, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv:1804.03999 [Preprint]. 2017. Available from: http://arxiv.org/abs/1706.05587.
12. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018.
13. Oktay O, Schlemper J, Folgoc LL, Lee M, Heinrich M, Misawa K, Mori K, McDonagh S, Hammerla NY, Kainz B, Glocker B, Rueckert D. Attention u-net: learning where to look for the pan-creas. arXiv:1804.03999 [Preprint]. 2018. Available from: http://arxiv.org/abs/1804.03999.
14. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Pro-ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018.
15. Chen Y, Kalantidis Y, Li J, Yan S, Feng J. a^2-nets: Double atten-tion networks. arXiv:1810.11579 [Preprint]. 2018. Available from: http://arxiv.org/abs/1810.11579.

16. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen Y-W, Wu J. Unet 3+: A full-scale connected Unet for medical image segmentation. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. p. 1055–9. https://doi.org/10.1109/ICASSP40776.2020.9053405.
17. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In: Wang Q, Shi Y, Suk H-I, Suzuki K, editors. Machine Learning in Medical Imaging. Cham: Springer International Publishing; 2017. p. 379–87.
18. Crum W, Camara O, Hill D. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans Med Imaging. 2006;25(11):1451–61. https://doi.org/10.1109/TMI.2006.880587.
19. Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). 2016. p. 565–71. https://doi.org/10.1109/3DV.2016.79.
20. Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.
22. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078 [Preprint]. 2014. Available from: http://arxiv.org/abs/1406.1078.
23. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. arXiv:1409.3215 [Preprint]. 2014. Avail-able from: http://arxiv.org/abs/1409.3215.
24. Luo P, Wang G, Lin L, Wang X. Deep dual learning for semantic image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017.
25. Zeiler MD, Taylor GW, Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. In: International Conference on Computer Vision, vol. 2011. 2011. p. 2018–25. https://doi.org/10.1109/ICCV.2011.6126474.
26. Pohlen T, Hermans A, Mathias M, Leibe B. Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
27. Peng C, Zhang X, Yu G, Luo G, Sun J. Large kernel matters –improve semantic segmentation by global convolutional network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
28. Amirul Islam M, Rochan M, Bruce NDB, Wang Y. Gated feed-back refinement network for dense image labeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recogni-tion (CVPR). 2017.
29. Hoang TM, Zhou, Fan JY. Image compression with encoder-decoder matched semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2020.
30. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M. U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recogn. 2020;106:107404. https://doi.org/10.1016/j.patcog.2020.107404
31. Wu Z, Su L, Huang Q. Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
32. Jaderberg M, Simonyan K, Zisserman A, Kavukcuoglu K. Spatial transformer networks. arXiv:1506.02025 [Preprint]. 2016. Available from: http://arxiv.org/abs/1506.02025.
33. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X. Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
34. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16. https://doi.org/10.1109/TPAMI.2015.2389824.
35. Lee Y, Park J. Centermask: Real-time anchor-free instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.
36. Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.
37. Orhan AE, Pitkow X. Skip connections eliminate singularities. arXiv:1701.09175 [Preprint]. 2018. Available from: http://arxiv.org/abs/1701.09175.
38. Cortes C, Gonzalvo X, Kuznetsov V, Mohri M, Yang S. AdaNet: Adaptive structural learning of artificial neural networks. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning, vol. 70 of Proceedings of Machine Learning Research. PMLR; 2017. p. 874–83. http://proceedings.mlr.press/v70/cortes17a.html.
39. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv:1802.06955 [Preprint]. 2018. Available from: http://arxiv.org/abs/1802.06955.