Computers in Biology and Medicine 134 (2021) 104449

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Hybrid dilation and attention residual U-Net for medical image segmentation

Zekun Wang [a], Yanni Zou [a,*], Peter X. Liu [b]

[a] The School of Information Engineering, Nanchang University, Jiangxi, 330031, China
[b] Department of Systems and Computer Engineering, Carleton University, Ottawa ON, K1S 5B6, Canada

ARTICLE INFO

Keywords:
Medical image segmentation
Convolutional neural network
Channel attention mechanism
Dilated convolution
Deep learning

ABSTRACT

Medical image segmentation is a typical task in medical image processing and critical foundation in medical image analysis. U-Net is well-liked in medical image segmentation, but it doesn't fully explore useful features of the channel and capitalize on the contextual information. Therefore, we present an improved U-Net with residual connections, adding a plug-and-play, very portable channel attention (CA) block and a hybrid dilated attention convolutional (HDAC) layer to perform medical image segmentation for different tasks accurately and effectively, and call it HDA-ResUNet, in which we fully utilize advantages of U-Net, attention mechanism and dilated convolution. In contrast to the simple copy splicing of U-Net in the skip connection, the channel attention block is inserted into the extracted feature map of the encoding path before decoding operation. Since this block is lightweight, we can apply it to multiple layers in the backbone network to optimize the channel effect of this layer's coding operation. In addition, the convolutional layer at the bottom of the "U"-shaped network is replaced by a hybrid dilated attention convolutional (HDAC) layer to fuse information from different sizes of receptive fields. The proposed HDA-ResUNet is evaluated on four datasets: liver and tumor segmentation (LiTS 2017), lung segmentation (Lung dataset), nuclear segmentation in microscope images (DSB 2018) and neuron structure segmentation (ISBI 2012). The dice global scores of liver and tumor segmentation (LiTS 2017) reach 0.949 and 0.799. The dice coefficients of lung segmentation and nuclear segmentation are 0.9797 and 0.9081 respectively, and the information theoretic score for the last one is 0.9703. The segmentation results are all more accurate than U-Net with fewer parameters, and the problem of slow convergence speed of U-Net on DBS 2018 is solved.

1. Introduction

Medical images play an indispensable and huge role in medical treatment and diagnosis. Medical image segmentation determines whether medical images can provide a reliable evidence in clinical diagnosis and treatment. It has extremely important research and application value in research fields such as disease diagnosis and analysis, medical research, and auxiliary surgery. Recently, deep learning methods, like other research fields of computer vision, have achieved excellent results in medical image segmentation, and are far ahead of traditional image segmentation approaches. At present, the method based on deep learning has been widely used in the field of image segmentation, including recurrent neural networks (RNN), convolutional neural networks (CNN), and fully convolutional networks (FCN).

RNN is increasingly used by researchers for segmentation tasks as it solves the problem that traditional neural networks cannot share location features from data. For example, Xie [1] segmented the muscle perimysium with spatial clockwork RNN. The network considers previous information from the row and column predecessors of the current patch. The bidirectional information from the right/bottom and left/top neighbors is combined and RNN is adopted four times in nonidentical directions. It directly considers 2D image, and can use local patches to represent global information about a whole image. However, spatial clockwork RNN contains more artificial components, the clock frequency must be manually set. Li et al. [2] adopted 3D CNN and LSTM recurrent units to extract the spatio-temporal information of multi-level for heart segmentation without pre-training, but the network is difficult to train and generalize. Compared with RNN, CNN has inherent superiority: the convolutional layer extracts features, the pooling layer corrects the attention direction, the fully connected layer explains the features. CNN and FCN are most used in medical image segmentation, and their segmentation accuracy is very high, which transcends the

* Corresponding author.
E-mail address: zouyanni@163.com (Y. Zou).

https://doi.org/10.1016/j.compbiomed.2021.104449
Received 29 January 2021; Received in revised form 11 April 2021; Accepted 23 April 2021
Available online 11 May 2021

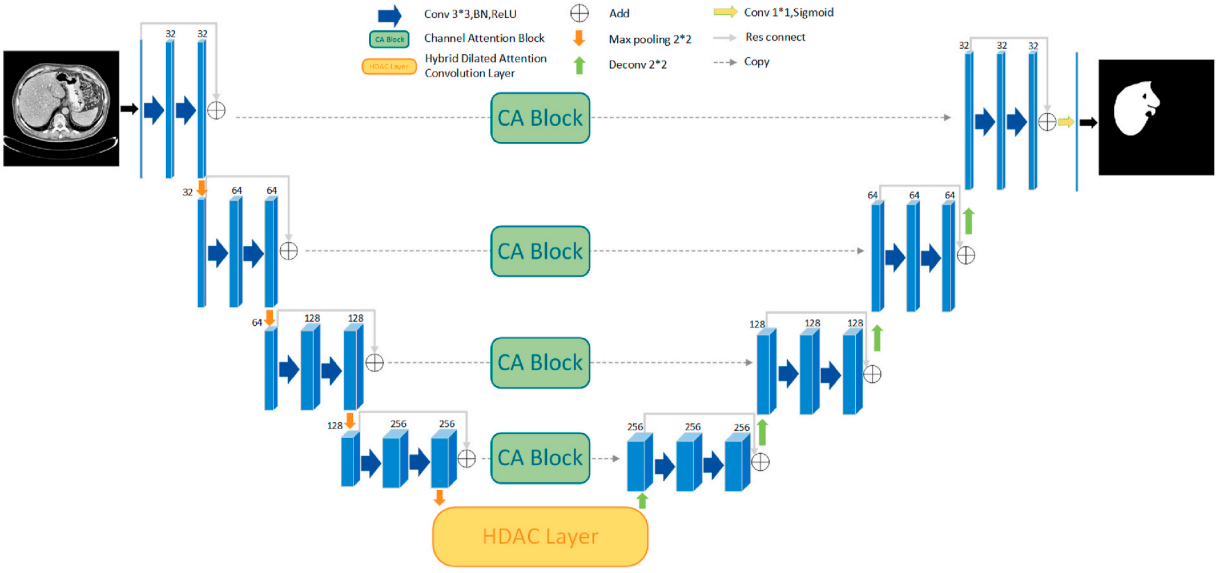

Fig. 1. HDA-ResUNet framework.

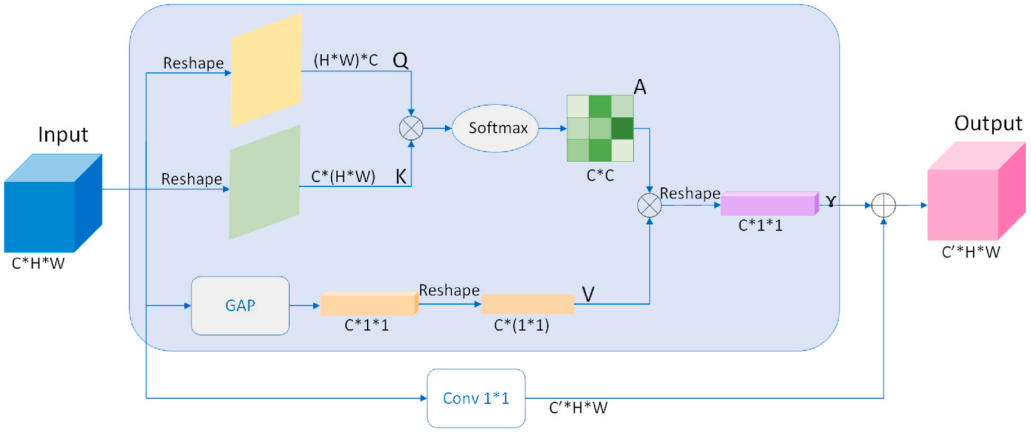

Fig. 2. The structure of the channel attention block.

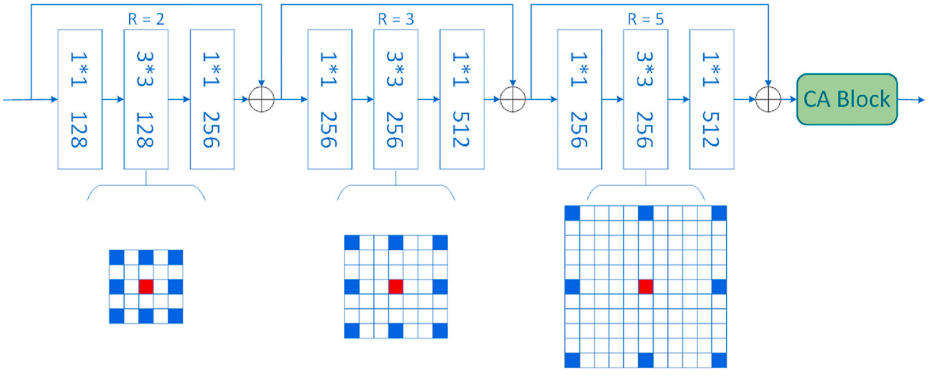

Fig. 3. The structure of the hybrid dilated attention convolutional layer.

traditional methods. They have excellent robustness and versatility, demonstrating the most advanced performance and the prospect of efficient automatic segmentation of soft tissue. In 2014, FCN was proposed by Long et al. [3], which superseded all the fully connected layers with convolutional layers. Not only the method can input data of any size, but also realize pixel-level image segmentation, which promotes the improvement of segmentation accuracy. On the basis of FCN, investigators have brought many outstanding fully convolutional neural networks, such as SegNet [4], DeepLab [5] and other networks [6–8], which have achieved excellent results in actual image segmentation. In 2017, Bi et al. [9] brought an improved FCN for automatic segmentation of skin lesions. Based on fully convolutional model, this method

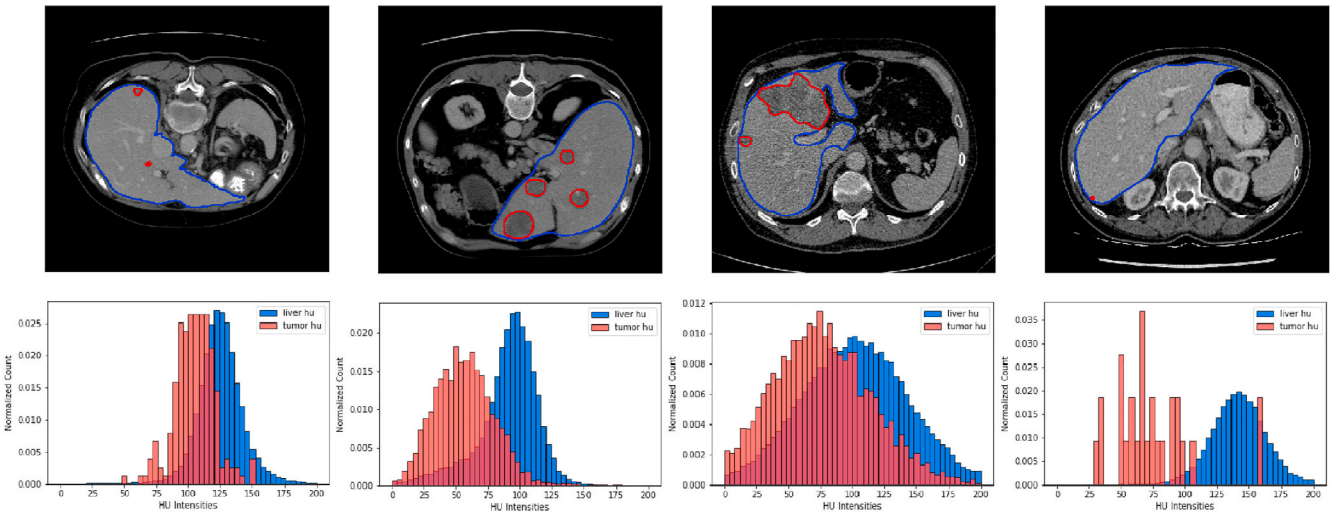

Fig. 4. Describe the density of HU-levels in the liver and tumor regions. The blue outlines represent segmented livers while the red ones represent the segmented tumors.

combines supplementary information from each segmentation stage to obtain more accurate lesion results.

The most worth mentioning is the U-Net introduced by Ronneberger et al. [10], which is an interesting innovation and this network has been preferably used in plenty of segmentation tasks. It is similar to the autoencoder [11] in that it consists of a contracting branch and an expanding branch. In addition, the network also adds skip connections to connect different features of the encoder and decoder, and models multi-scale image-to-image transformations with different resolutions. It can work with limited training samples, which solves the matter that the scarcity of medical image segmentation data and the imbalance of categories. So far, investigators have already brought plenty of improvement schemes based on U-Net [12–20]. For example, by adding residual connections to reduce the impact of gradients, good results have been achieved in segmentation fields, such as retinal vessels [15], brain tumors [16], and optic nerve head [17]. But for medical segmentation, it is necessary to focus on essential features. The attention mechanism helps draw attention to what we want. It is commonly found in text processing, speech recognition, image processing and other fields. Its main feature is that it can dynamically assign the input weights of neurons to selectively focus on the most critical part of the information. Vaswani et al. [21] first proposed a self-attention mechanism to draw the global dependence and implemented it to machine translation. In the meantime, more and more attention modules are showing up. Wang et al. [22] introduced self-attention to computer vision, and explored the effectiveness of non-local operations of videos and images in the space-time dimension. Zhang et al. [18], Bhatkalkar et al. [19], Oktay et al. [20] applied the attention gate to medical image segmentation. Such networks are effective in fusing multi-level features, but do not fully utilize the contextual information and discover useful features of the channel.

In this paper, we propose HDA-ResUNet, an improved version of U-Net. First, we add residual connections to each layer of the left and right branches to help network training. Meanwhile, unlike U-Net, which extracts feature mappings in skip connections and then directly replicates them for splicing to the decoder, a self-attentive mechanism is introduced to capture feature dependencies in the channel dimension while ensuring effective fusion of multi-level features. It represents feature dependencies and channel correlation on channel dimensions by assigning the specific attention weight to channels. By including channel attention in the skip connection, it helps to reduce the effect of noise and gives more attention to essential regions. In addition, we improve the bottom structure of U-Net by proposing a hybrid dilated attention convolutional (HDAC) layer, which effectively increases the receptive field, aggregates global information, and reduces the impact of learned redundant features to obtain a better segmentation effect. We evaluate four different applications: liver and tumor segmentation (LiTS 2017), lung segmentation (Lung dataset), nuclear segmentation in microscope images (DSB 2018) and neuron structure segmentation (ISBI 2012). Experimental results show that HDA-ResUNet achieves higher performance than U-Net with fewer parameters.

The rest of this paper is organized as follows. Details of the proposed method are described in Section 2. In Section 3, the datasets, experiments and results are introduced in detail. Discussion and future work are provided in Section 4. Conclusion is summarized in Section 5.

2. Methodology

Inspired by U-Net, DeepLab and SENet [23], we propose HDA-ResUNet, as shown in Fig. 1. The network takes advantage of the long short skip connection to fully combine the low-level and high-level information, and replaces the bottom of the "U"-shaped network with a hybrid dilated attention convolutional layer to acquire multi-scale objects and context information, and uses the channel attention block to focus more on channels which have ample information, while restraining those inconsequential channels. In the next part, we will detail the different parts of the network.

2.1. HDA-ResUNet framework

Fig. 1 shows the basic description of the HDA-ResUNet framework. The network structure is composed of an encoding branch (left one) and a decoding branch (right one). The encoding branch repeatedly applies the basic residual block with two consecutive convolutional layers (same padding). Here, each convolutional layer is followed by a batch normalization layer and a ReLU nonlinear layer. Each basic residual block is followed by a 2×2 max-pooling operation for down-sampling $= 0.30$em. At each down-sampling step we double the number of feature channels, and the default initial number of channels is $32 = 0.30$em. Then, the hybrid dilated attention convolutional layer at the bottom summarizes the global information and generates the output of the encoder. Correspondingly, the same number of up-sampling processes is performed in the decoding branch to restore the space size of the segmented output. Each up-sampling is implemented by a 2×2 transposed convolution and the number of feature channels is halved. In order to assist the decoding process, skip connections are made to copy features from the encoder to the decoder after passing through the

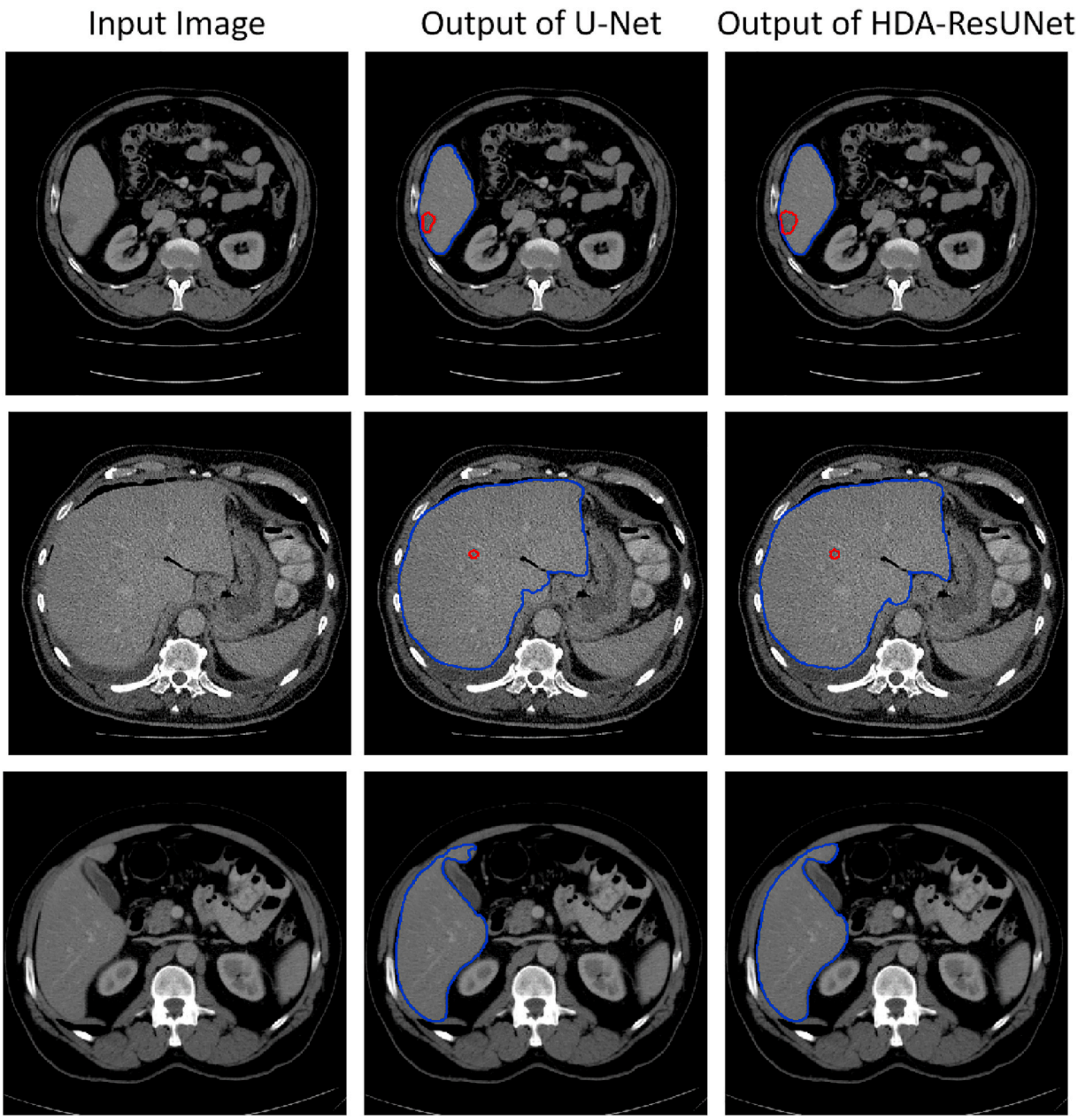

Fig. 5. Comparison of segmentation results on LiTS 2017. From the first column to the third one, they show in order: the input sample, the prediction of U-Net, the prediction of HDA-ResUNet. The blue outlines represent segmented livers while the red ones represent the segmented tumors.

Table 1
Performance comparison between the proposed network and other networks on LiTS 2017. SENet follows the same structure as HDA-ResUNet, with SE blocks (defined in Ref. [23]) instead of CA blocks. Dice global score (DG) is calculated by combining all CT volumes into one, and dice per case score (DC) is the mean dice score of each volume.

Methods	Liver		Tumor	
	DG	DC	DG	DC
U-Net [10]	0.941	0.935	0.779	0.620
Attention U-Net [20]	0.948	0.944	0.799	0.625
HTTU-Net [32]	0.947	0.943	0.767	0.623
RA-UNet [33]	0.963	0.961	0.795	0.595
DRUNET [17]	0.929	0.920	–	–
DeepLab [5]	0.945	0.941	0.781	0.595
SENet [23]	0.946	0.943	0.786	0.652
HDA-ResUNet	**0.949**	**0.944**	**0.799**	**0.653**

channel attention block, and then basic residual blocks with two consecutive convolutional layers (same padding) are employed for feature extraction. Similarly, each convolutional layer is followed by a batch normalization layer and a ReLU nonlinear layer. Differently, the encoded features are merged with decoded features by summation, rather than simple copy concatenation used in U-Net. The intuitive way to combine features from the encoder and the decoder is concatenation, which provides two sources of inputs to the up-sampling operation. While, using the additive approach to summarize features has two advantages [24]. Firstly, the number of feature maps is not expanded by summation, thus training parameters in the next layer are reduced accordingly. Next, skip connections with summation may be regarded as long-range residual connections, which can effectively promote the model training.

In this network structure, a large number of batch normalization layers [25] are used. BN layers assist the neural network to enhance stability and speed up the training process, which standardize the inputs

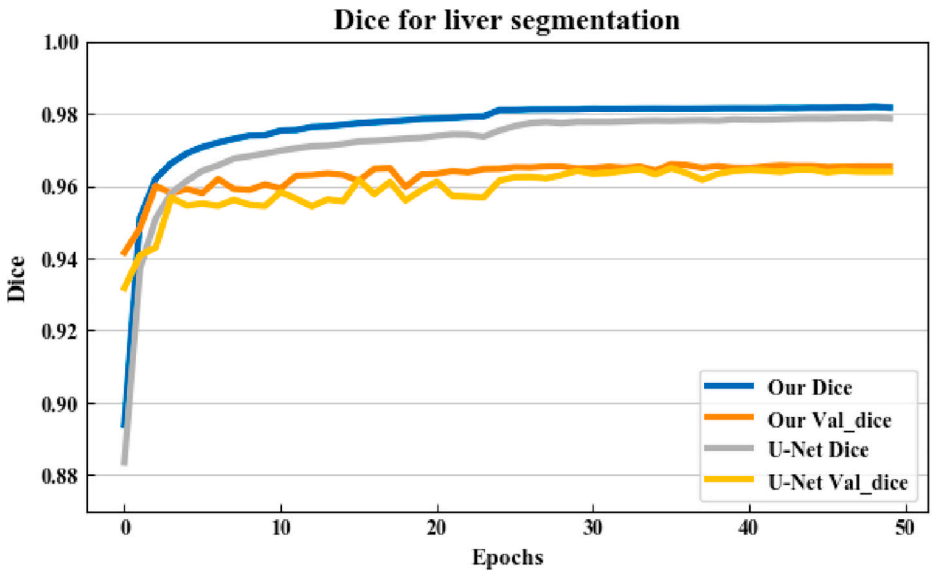

Fig. 6. Dice coefficients of the HDA-ResUNet and U-Net for liver segmentation.

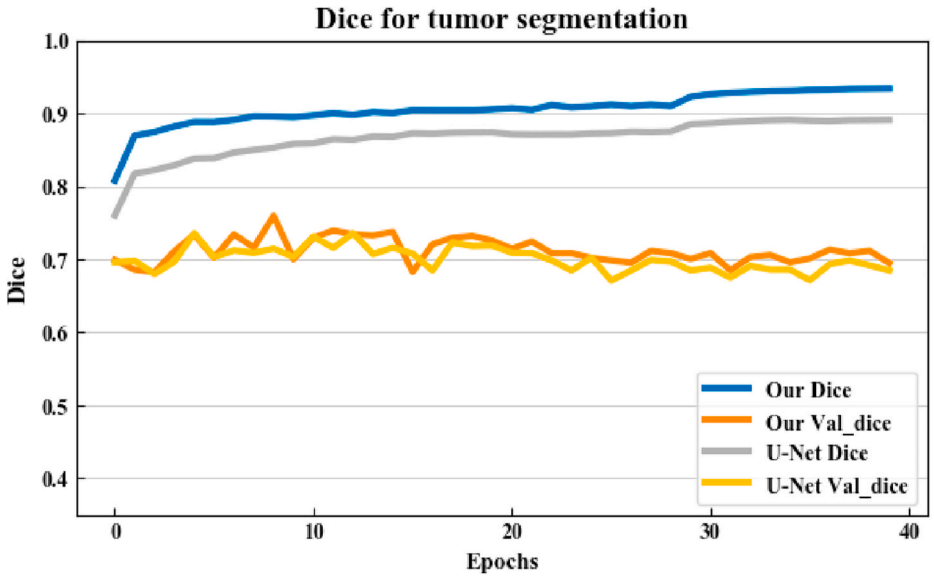

Fig. 7. Dice coefficients of the HDA-ResUNet and U-Net for liver tumor segmentation.

of layers in the network. In addition, the result of the model is ameliorated due to the moderate regularization effect in certain circumstances. Moreover, our network does not use the full connection layer, which ensures that the segmentation results are predictions based on contextual information, and also reduces model parameters. In the encoding process, the encoder acquires high-level feature by constantly using a stack of blocks, which consist of convolutional layers and max-pooling layers, and the receptive fields of the convolutional kernels are multiplied in each down-sampling process. In the decoding process, the feature output from the previous layer is upsampled and fused with features from the skip connection, and then the reconstruction of input information is accomplished by continuous convolutional layers. By continuous up-sampling, the final output size is the same as the input image size.

In addition, we have added short skip connections, that is, residual structure [26]. The residual structure makes the network optimization easier, speeds up model convergence, and enhances accuracy by increasing the model depth. The appropriate combination of long and short connections helps the network to extract features at different levels and enhance the ability of expression, while complementing semantic information at the high level and refining segmentation outlines at the low level.

2.2. *Channel attention block*

Each channel corresponds to a specific semantic response. Different channels have different contributions to acquire useful feature information. We hope to improve the representation ability of the network by modeling the dependencies of each channel, and adjust the features channel by channel, so that the network can learn to selectively strengthen the features containing useful information and suppress the features containing useless information. Adding the channel attention (CA) block for each skip connection to eliminate redundant information that is irrelevant during skip connections. By leveraging the interdependence of channel maps, we stress the interdependent feature maps and refine the particular feature representation. Consequently, we

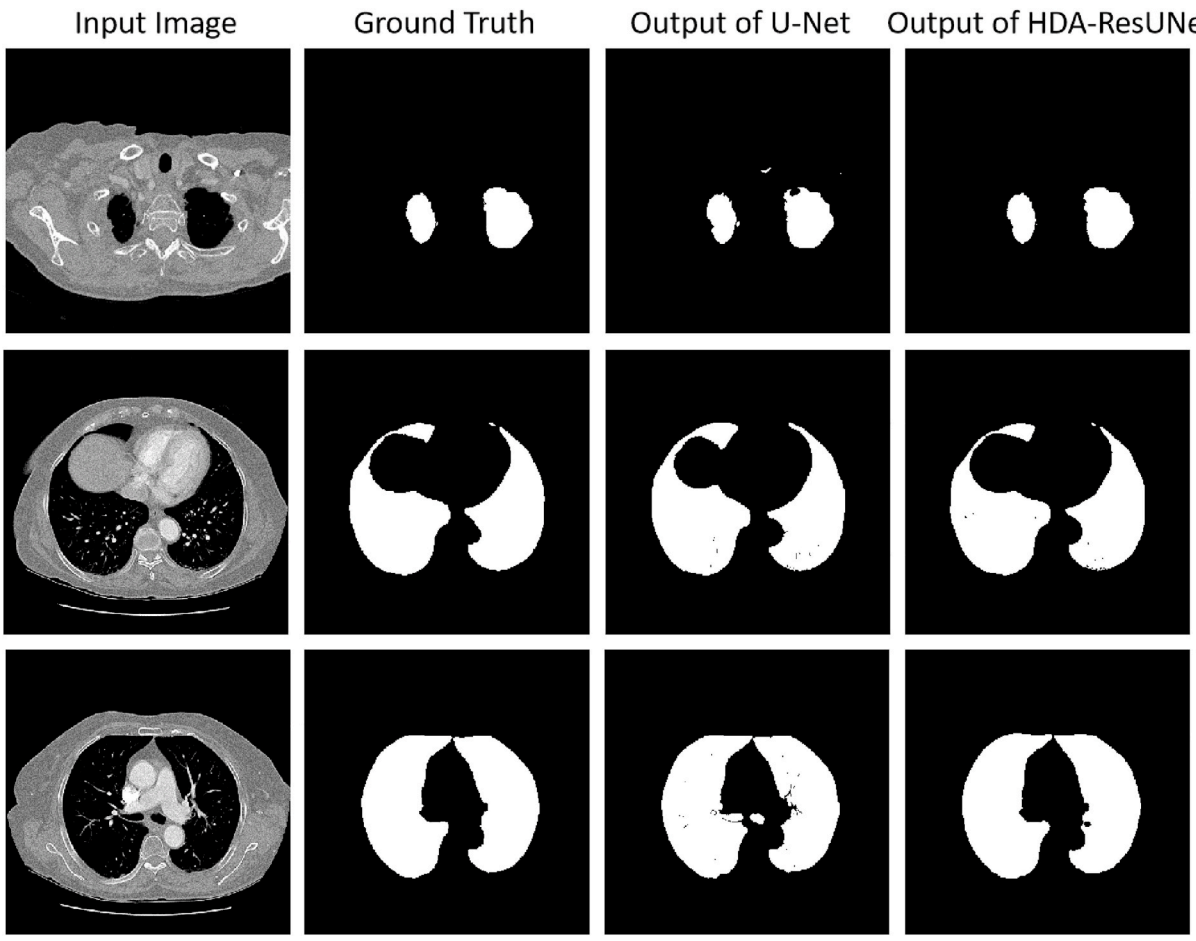

Fig. 8. Comparison of segmentation results on the Lung dataset. From the first column to the fourth one, they show in order: the input sample, the ground truth, the prediction of U-Net, the prediction of HDA-ResUNet.

Table 2
Comparison of the performance of the proposed network with other networks on the Lung dataset.

Methods	Dice	JS	Acc	SE	SPE
U-Net [10]	0.9543	0.9126	0.9856	0.9213	0.9881
Attention U-Net [20]	0.9267	0.8635	0.9753	0.9591	0.9784
HTTU-Net [32]	0.9679	0.9379	0.9895	0.9668	0.9940
RA-UNet [33]	0.9655	0.9333	0.9887	0.9682	0.9927
DRUNET [17]	0.9776	0.9563	0.9927	0.9709	0.9970
DeepLab [5]	0.9734	0.9482	0.9914	0.9654	0.9964
SENet [23]	0.9438	0.8937	0.9823	0.9098	0.9965
HDA-ResUNet	**0.9797**	**0.9603**	**0.9934**	**0.9743**	**0.9971**

construct a channel attention (CA) block to explicitly simulate the interdependence of channels. Fig. 2 shows the module composition and structure in detail.

First, we reshape the input feature map $I \in \mathbb{R}^{C*H*W}$, reshaping I into $K \in \mathbb{R}^{C*(H*W)}$, $Q \in \mathbb{R}^{(H*W)*C}$ respectively, and then divide the product of matrix K and Q by the factor $\sqrt{C}$. Finally, we apply a *Soft*max layer to get the channel attention map $A \in \mathbb{R}^{C*C}$.

$$a_{ji} = Softmax\left(\frac{f(I_i, I_j)}{\sqrt{C}}\right) \tag{1}$$

Where, a_{ji} indicates the i^{th} channel's influence on j^{th} channel, and the function f is used to calculate the relationship between i and all j. We use global average pooling (GAP) to acquire the channel statistics of each channel of the original feature map $I \in \mathbb{R}^{C*H*W}$ compressed into global spatial information. As shown in Equation (2), global average pooling compresses global information and obtains an attention vector to achieve feature dimensionality reduction and high-level semantic information extraction. By encoding this vector, salient features can be preserved. We execute a matrix multiplication between A and V, and transform the result into $\mathbb{R}^{C*1*1}$, then multiply it by the proportional parameter γ, and use I after 1×1 convolution for summation to obtain the final output $O \in \mathbb{R}^{C*H*W}$.

$$g(I_k) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} I_k(i, j) \tag{2}$$

Where, g stands for global average pooling, and $k = 1, 2, \ldots, c, I = [i_1, i_2, \ldots, i_c]$.

$$O_j = \gamma \sum_{i=1}^{C} \left(a_{ji} \cdot g(I_k)\right) + \omega_\theta I_j + b_\theta \tag{3}$$

Where γ starts from 0 and learns the weight during training, gradually. ω_θ refers to the weight of the 1×1 convolution. b_θ is the bias. From Equation (3), the final output is a weighted sum between features that come from self-attention with GAP, and feature maps obtained by convolution. And we utilize the spatial information of all corresponding locations to model the channel correlation and also employ the global pooling to explore the channel relationship.

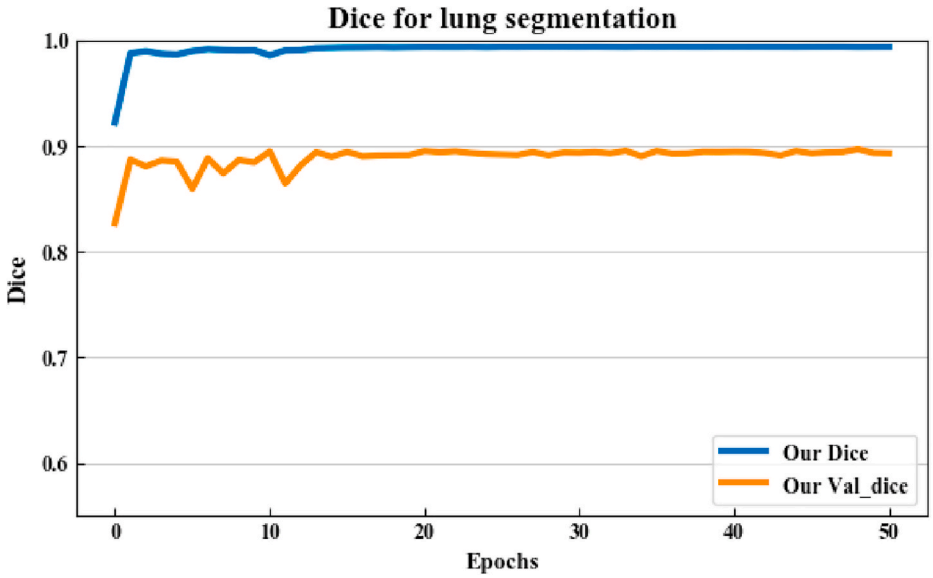

Fig. 9. Training and validation dice coefficients of HDA-ResUNet for lung segmentation.

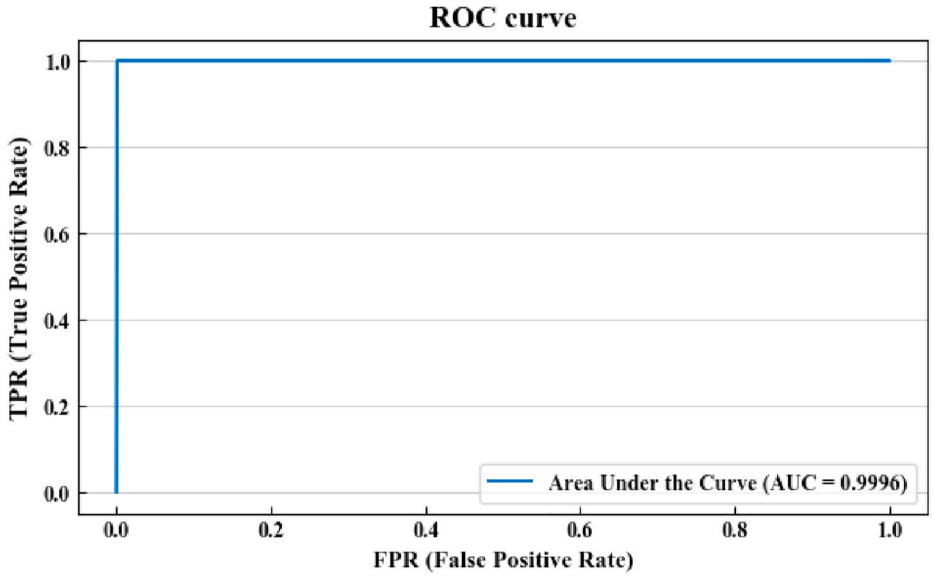

Fig. 10. ROC curve of HDA-ResUNet on the Lung dataset.

2.3. Hybrid Dilated attention convolutional (HDAC) layer

The structure of hybrid dilated attention convolutional (HDAC) layer is shown in Fig. 3, which is composed of 3 bottleneck dilated modules with different dilated rates and 1 channel attention module. Among them, the bottleneck dilated module is composed of convolutional layers of 1×1, 3×3, 1×1 (the convolution of 3×3 is the dilated convolution), and each bottleneck dilated module also has a residual connection. Compared with traditional convolution, the remarkable distinction of dilated convolution is the arbitrary expansion of the receptive field without adding extra parameters. In a two-dimensional image, dilated convolution can be defined as Equation (4):

$$v[i] = \sum_{k=1}^{K} u[i + d \cdot k] f[k] \tag{4}$$

where u[i] is the input signal, v[i] is the output signal, f[k] is the filter with a convolution kernel size of k, and d is the dilated rate of the dilated convolution. $d = 1$ is the standard convolution.

According to Equation (4), for the kernel of convolution is k, derived dilated filter with size k', where $k' = k + (k-1)\cdot(r-1)$. Large receptive fields can better segment large objects. Although only using the information obtained by convolution with large dilated rate may be effective for the segmentation of some large objects, it is not helpful for small objects. The key to using dilated convolution is how to handle different sizes of objects simultaneously. Arbitrary use of dilated convolution is easy to produce checkerboard artifacts. Therefore, we adopt the hybrid dilated convolution brought by Wang et al. [27] to reduce the influence of checkerboard artifacts. The dilated convolution rate of cascade is designed to be 2, 3 and 5, so as to meet the conditions and objectives of Equation (5).

$$M_i = Max[d_i, M_{i+1} - 2d_i, M_{i+1} - 2(M_{i+1} - d_i)] \tag{5}$$

Where d_i is the dilated rate of the i layer, and M_i refers to the maximum dilated rate in the i layer. Assuming that the convolution kernel size is $k \times k$, the target is $M_2 \leq k$.

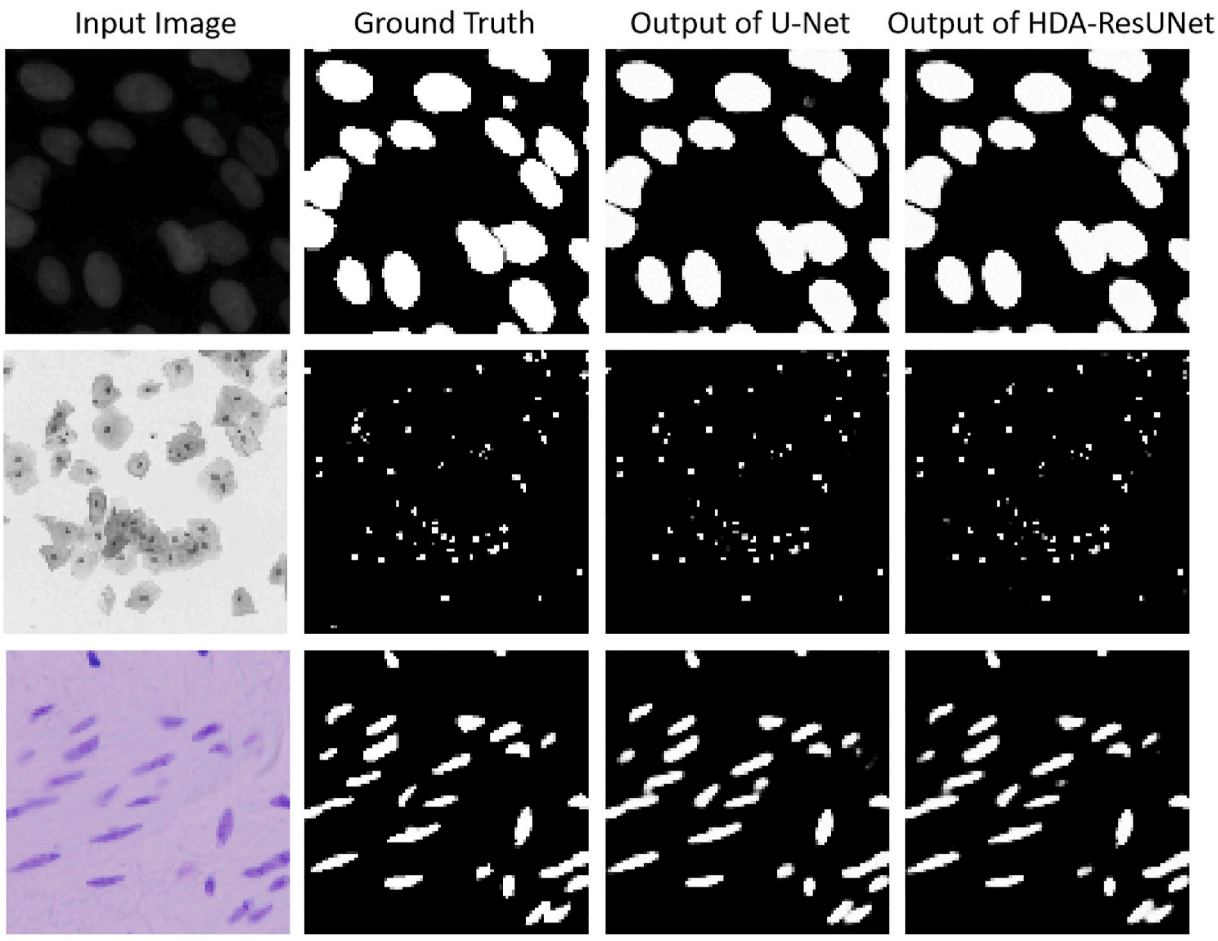

Fig. 11. Comparison of segmentation results on DSB 2018. From the first column to the fourth one, they show in order: the input sample, the ground truth, the prediction of U-Net, the prediction of HDA-ResUNet.

Table 3
Performance comparison between the proposed network and other networks on DSB 2018.

Methods	Dice	JS	Acc	Pre	AUC
U-Net [10]	0.9019	0.8287	0.9728	0.8220	0.9554
Attention U-Net [20]	0.9034	0.8308	0.9734	0.8311	0.9534
HTTU-Net [32]	0.9066	0.8361	0.9741	0.8267	0.9579
RA-UNet [33]	0.8697	0.7750	0.9619	0.8305	0.9240
DRUNET [17]	0.9031	0.8318	0.9734	0.8261	0.9556
DeepLab [5]	0.6005	0.4369	0.8587	0.4847	0.8085
SENet [23]	0.9068	0.8348	0.9747	0.8313	0.9564
HDA-ResUNet	**0.9081**	**0.8370**	**0.9745**	**0.8341**	**0.9557**

By cascading multiple bottleneck dilated modules with different dilated rates, the fusion of receptive field information of different sizes is realized, multi-scale context information is fully extracted, and the number of parameters is effectively reduced. We add residual connections for each bottleneck dilated module, which is conducive to network optimization. At the same time, we use the channel attention (CA) block mentioned above to perform nonlinear fusion of the information of each channel. The nonlinear function is applied to represent the relationship between the context information of different channels, and then the weights are assigned to the multi-scale context information to facilitate extraction of crucial features.

3. Experiments and results

We evaluated HDA-ResUNet on LiTS 2017, a public benchmark dataset of lung segmentation, DSB 2018 and ISBI 2012. LiTS 2017 is a LiTS competition dataset jointly organized by MICCAI 2017 and ISBI 2017. It is used for the challenges of liver tumor segmentation. The second dataset is obtained from the kaggle challenge and used for lung segmentation of CT images. DSB 2018 is a dataset for nuclear segmentation of microscope images. ISBI 2012 is a dataset for segmenting the neuronal structures recorded by electron microscopy. CT images are grayscale images, so input channels of CT images were 1, and microscope images are color images, so input channels of those were 3. Experimental results show that the proposed method has good performance with fewer parameters. The code was implemented using Pytorch. The model was trained on the CentOS operating system, with a 12 GB Nvidia TITAN V GPU. The training of the four datasets all used the ADAM [28] optimizer, the initial learning rate was 0.001, and the gradient descent algorithm was implemented. We considered several performance indicators to perform experimental comparisons, including dice coefficient (Dice), jaccard similarity (JS), accuracy (Acc), precision (Pre), sensitivity (SE), specificity (SPE), area under receiver operating characteristic curve (AUC). The related formulas are described in Equations 6–11. In particular, according to the evaluation requirements of ISBI challenge in 2012, we adopt rand score and information theoretic score (Info score) [29]to evaluate the neuron segmentation performance.

$$\mathrm{Dice} = \frac{2|S \cap G|}{|S| + |G|} \tag{6}$$

$$\mathrm{JS} = \frac{|S \cap G|}{|S \cup G|} \tag{7}$$

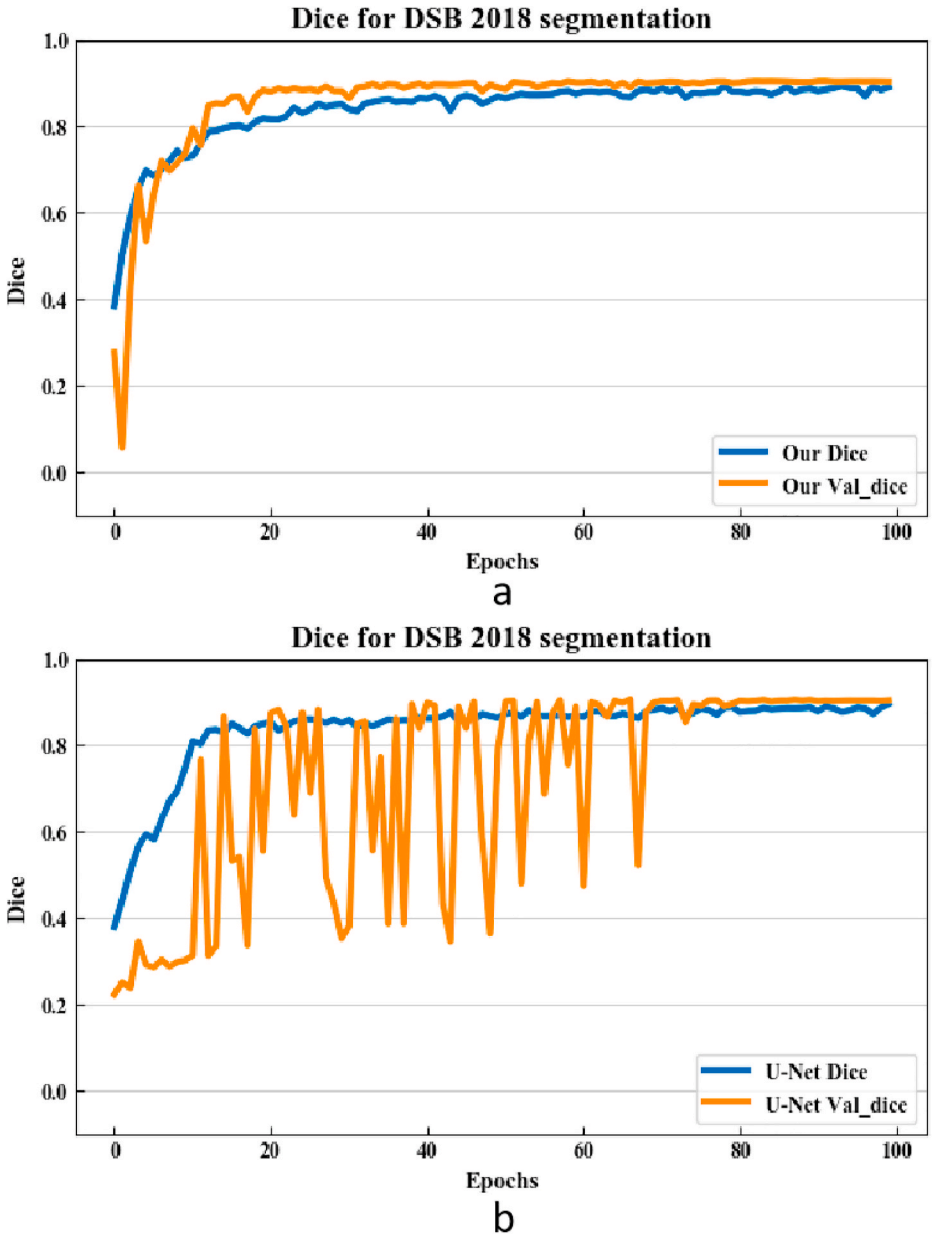

Fig. 12. Training and validation of the dice coefficients on DSB 2018.

Where, S and G respectively represent segmentation results and manual labelings.

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{8}$$

$$\mathrm{Pre} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

$$\mathrm{SPE} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{11}$$

Where, TP, TN, FP, and FN respectively represent true positives, true negatives, false positives, and false negatives.

3.1. Performance on LiTS 2017

LiTS 2017 [30] is a dataset of the LiTS competition co-organized by MICCAI and ISBI in 2017. The dataset provides 131 publicly downloadable manually annotated abdominal enhancement CT images as the training set and 70 images of unavailable ground truth as the test set. These data are obtained from different patients at seven different hospitals and research institutions around the world. The ground truth is annotated by three different radiologist to locate the area of the liver and liver tumor. Each image set consists of a abdominal CT enhancement image and a manually labeled image of the same size, with varying slice spacing from 0.45 mm to 6.0 mm, and a wide range of cross-sectional resolution from 0.55 mm to 1.0 mm. Of these, 118 samples were used for training, 13 samples for validation. For liver segmentation, we resampled all the section spacing to 1 mm, input size was 352×352, the window width and window level were 450 and -25 respectively, and only the sections containing liver were retained as the training data. In order to ensure generalization, 30% of the sections containing liver were expanded to include the adjacent sections without liver. As shown in

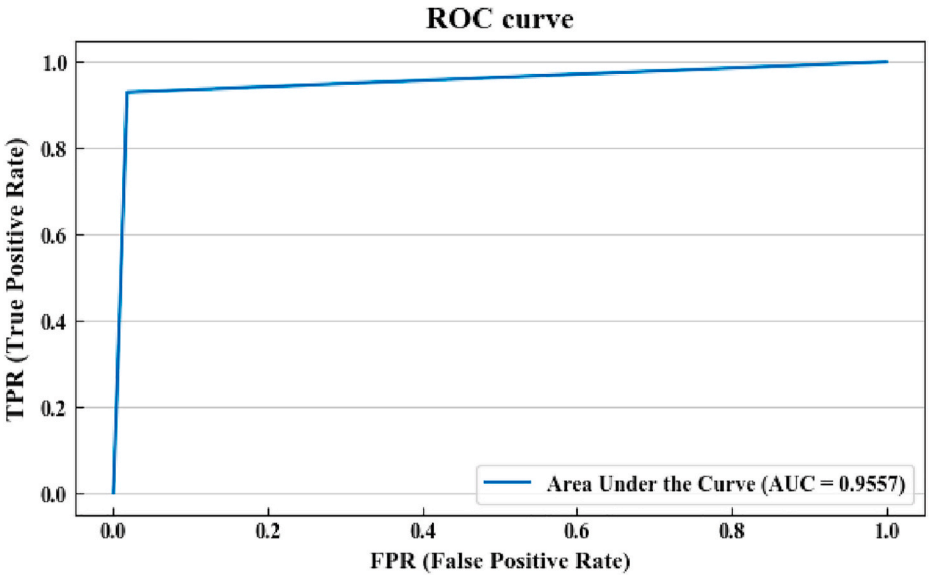

Fig. 13. ROC curve of HDA-ResUNet on DSB 2018.

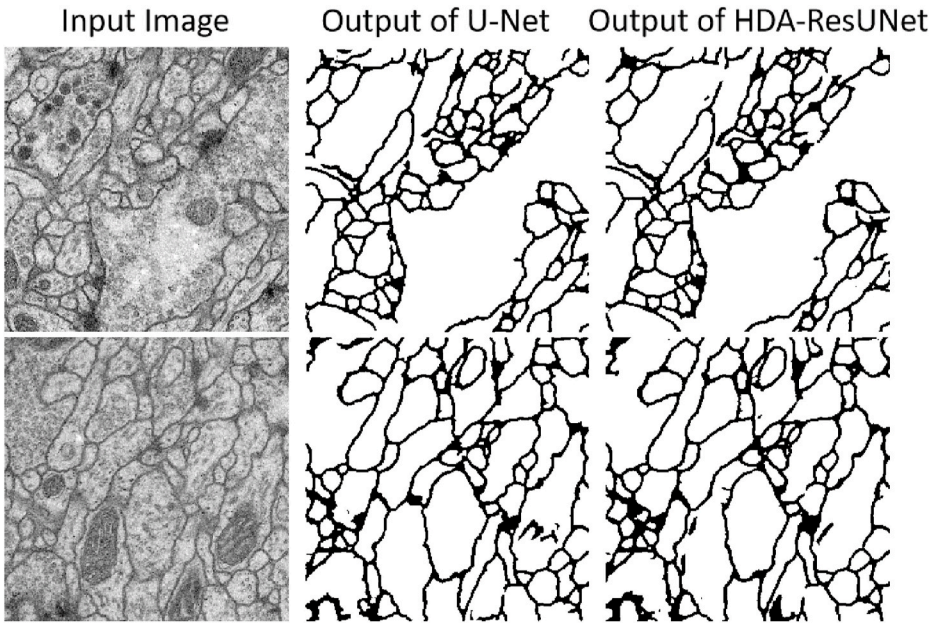

Fig. 14. Comparison of segmentation results on ISBI 2012. From the first column to the third one, they show in order: the input sample, the prediction of U-Net, the prediction of HDA-ResUNet.(The ground truth for ISBI 2012 is not given.)

Table 4
Performance comparison between the proposed network and other networks on ISBI 2012.

Methods	Rand score	Info score
U-Net [10]	0.8789	0.9567
Attention U-Net [20]	0.5744	0.7791
HTTU-Net [32]	0.8648	0.9545
RA-UNet [33]	0.6943	0.8833
DRUNET [17]	0.8501	0.9437
DeepLab [5]	0.1445	–
SENet [23]	0.8805	0.9606
HDA-ResUNet	**0.9215**	**0.9703**

Table 5
Comparison of the number of parameters between the proposed model and other models.

Methods	Initial channel	Parameters(Million)
RA-UNet [33]	16	11
DRUNET [17]	16	4
U-Net [10]	32	8
HTTU-Net [32]	64	53
U-Net [10]	64	31
Attention U-Net [20]	64	35
HDA-ResUNet	32	6
HDA-ResUNet	64	16

Table 6
Ablation study by comparing segmentation performance between different models on LiTS 2017. Dice global score (DG) is calculated by combining all CT volumes into one, and dice per case score (DC) is the mean dice score of each volume.

Methods	Liver		Tumor	
	DG	DC	DG	DC
Model 1	0.946	0.941	0.776	0.616
Model 2	0.940	0.935	0.790	0.623
Model 3	0.939	0.935	0.765	0.629
HDA-ResUNet	**0.949**	**0.944**	**0.799**	**0.653**

Table 7
Ablation study by comparing segmentation performance between different models on Lung dataset.

Methods	Dice	JS	Acc
Model 1	0.9648	0.9320	0.9884
Model 2	0.9671	0.9364	0.9893
Model 3	0.9748	0.9508	0.9917
HDA-ResUNet	**0.9797**	**0.9603**	**0.9934**

Table 8
Ablation study by comparing segmentation performance between different models on DSB 2018.

Methods	Dice	JS	Acc
Model 1	0.9046	0.8326	0.9737
Model 2	0.8969	0.8181	0.9714
Model 3	0.8975	0.8182	0.9707
HDA-ResUNet	**0.9081**	**0.8370**	**0.9745**

Table 9
Ablation study by comparing segmentation performance between different models on ISBI 2012.

Methods	Rand score	Info score
Model 1	0.9103	0.9654
Model 2	0.8995	0.9675
Model 3	0.8940	0.9664
HDA-ResUNet	**0.9215**	**0.9703**

Fig. 4, it can be seen that the heterogeneity of liver and tumor varies widely among patients, so we applied the histogram equalization [31]. Liver segmentation cost approximately 120 h to train for 50 epochs with a pixel-weighted binary cross-entropy loss function. For tumor segmentation, since tumors are mostly small and harder to train, we chose the trained liver model as the pre-trained model for tumors and used the original resolution as the input to prevent artifacts, while setting the initial number of channels to 64, the learning rate to 0.00003. It took approximately 50 h to train for 40 epochs with the same loss function as liver segmentation. Fig. 5 shows some accurate segmentation examples from the experimental results of the HDA-ResUNet and U-Net.

Table 1 presents the quantitative results acquired on LiTS 2017 by different methods. HDA-ResUNet has achieved a great improvement. The dice coefficients of U-Net and the proposed network for training and validation on LiTS 2017 are shown in Figs. 6 and 7. As we can see that, the proposed HDA-ResUNet has higher dice coefficients for both training set and validation set than the classical U-Net on the LiTS 2017 dataset. Thus, we can learn that the learning ability and generalization ability of HDA-ResUNet are stronger than that of U-Net.

3.2. Performance on the lung dataset

The Lung dataset is from the Kaggle challenge. Lung segmentation is quite significant for the diagnosis of lung related diseases and can be utilized to lung cancer segmentation and lung nodule detection. That comprises 2D and 3D CT images, as well as images of the corresponding lung distribution [34]. We used 60% of the accessible samples for training, 20% for validation, and 20% for testing. Each image resolution was 512 × 512. The cutoff threshold was [-512, 512] to remove irrelevant factors such as bones and blood vessels. It cost approximately 5 h to train for 50 epochs with a binary cross-entropy loss function. Fig. 8 shows several accurate segmentation examples from the experimental results of the HDA-ResUNet and U-Net.

Table 2 indicates the quantitative results of the diverse methods on the dataset of lung segmentation, and excellent results were obtained on each of the evaluation metrics. The training and validation dice coefficients of HDA-ResUNet for lung segmentation are shown in Fig. 9. The convergence of the network on Lung dataset is fast (after 20 epochs). To present the overall performance of HDA-ResUNet on Lung dataset, Fig. 10 shows the ROC curve.

3.3. Performance on DSB 2018

DSB 2018 [35] is a dataset used for nuclear segmentation in microscope images. The dataset contains 670 nuclear images. The images are acquired under diverse circumstances with different cell type, imaging methods (such as white light illumination and fluorescent illumination) and magnification. This dataset is designed to challenge the ability of algorithms to summarize these changes. Among them, 536 samples were used for training, 67 samples were used for validation, and 67 samples were used for testing. Using the same preprocessing as [36], we integrated all of nuclei masks belonging to the same image (each mask contains a core, and the masks do not overlap, that is, no pixels and data belonging to two masks) into one mask as the mask of input image. And then each image was adjusted to 96 × 96. The training data contains the original image and the corresponding mask annotation. It cost approximately 20 min to train for 100 epochs with dice loss function. Fig. 11 shows several accurate segmentation examples from the experimental results of the proposed network and U-Net.

Table 3 shows the quantitative results acquired by various approaches on the DSB 2018 dataset. The results of the proposed HDA-ResUNet were better than those of U-Net. Most of the evaluation metrics gained higher results. Fig. 12 shows the training and validation dice coefficients of HDA-ResUNet and U-Net on DSB 2018 respectively. According to Fig. 12, we can intuitively see that our proposed network can reach convergence about 20 epochs, while the classical U-Net needs about 80 epochs, which saves 75% of the time. However, some of the dice in the validation set are higher than that in the corresponding training set. The primary factors causing this phenomenon are as follows: 1) the validation is tested with a model that has already been trained for one epoch at least. This model has learned a lot through massive training. 2) enhancement (rotation, flip etc.) enriches the training set, makes the data more diversified, and makes learning more difficult, but validation and testing are not performed data enhancement. To present the overall performance of HDA-ResUNet on the DSB 2018 dataset, Fig. 13 shows the ROC curve.

3.4. Performance on ISBI 2012

ISBI 2012 [37] is a dataset that dissects the neuronal structures recorded by the electron microscope. This dataset is provided by EM Segmentation Challenge and is still receiving new contributions. It contains 30 training slices of 512 × 512px drosophilus ventral nerve cords, as well as the same number of test slices with the same resolution, and provides segmentation ground truth for each training image. However, we can't access to the ground truth of the test. An forecast result can be sent to the organizers for evaluation. By calculating rand score and information theoretic score to complete the assessment. It took approximately 4 min to train for 50 epochs with dice loss function.

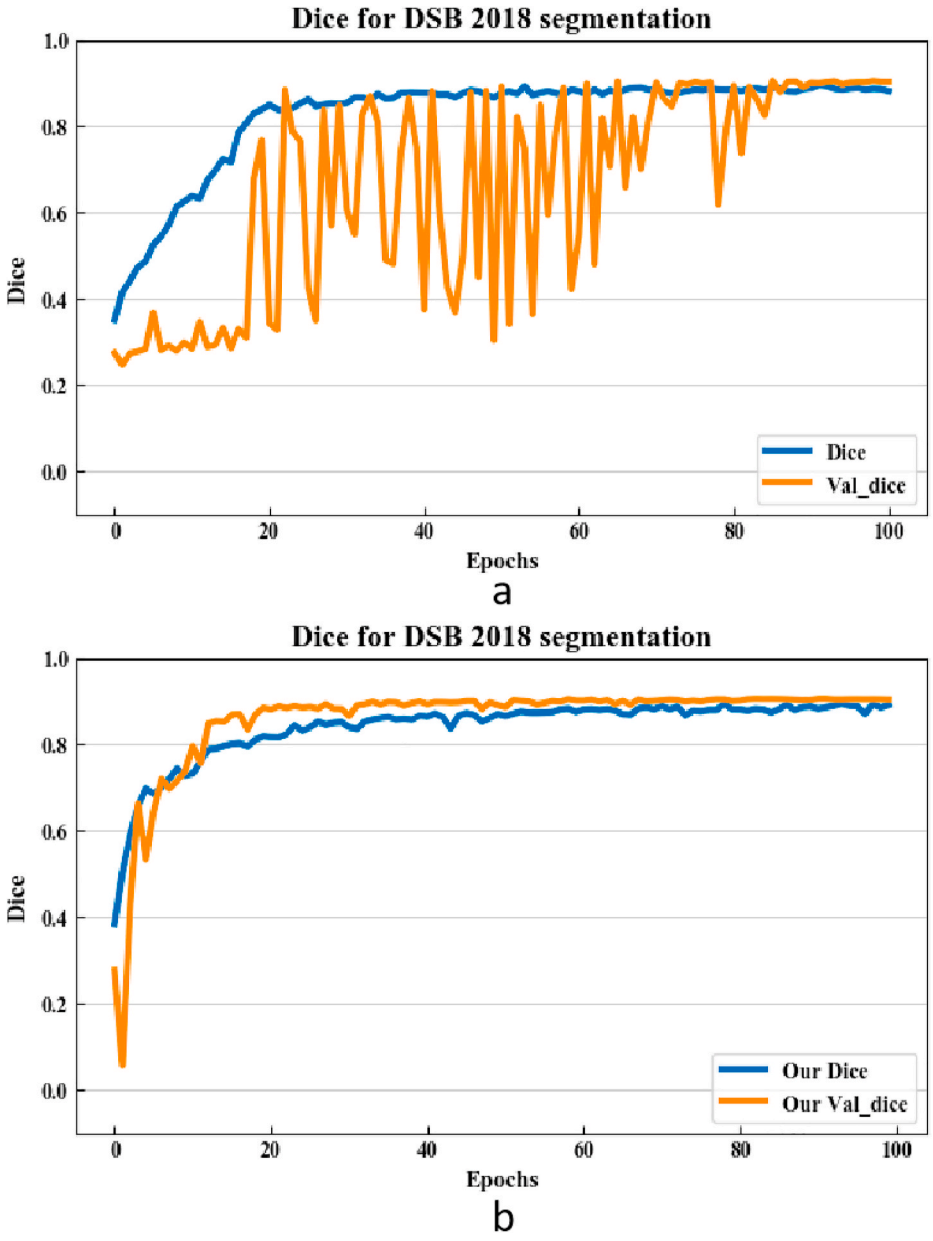

Fig. 15. Training and validation dice coefficients of HDA-ResUNet (a) without and (b) with residual connection on DSB 2018.

Examples of the dataset and the detailed results are shown in Fig. 14. Table 4 compares the quantitative results of the proposed HDA-ResUNet with those of other approaches. It is obvious that HDA-ResUNet is superior to other methods.

To further investigate the efficiency of our proposed HDA-ResUNet, we introduce a comparison of model parameters in Table 5. All models are calculated under the same quantitative standard. With the same initial number of channels, the HDA-ResUNet parameters are reduced by 25% at least compared to U-Net. With less number of parameters, better performance is obtained.

3.5. Ablation studies of different modules

The network we proposed has made some improvements on the basis of the original U-Net. To explore and confirm the effectiveness of various modules we designed, comparison experiments are conducted respectively. Concretely, we compare the following models:

Model 1 is a "U"-structured model. Down-sampling strategy is max-pooling, and up-sampling is implemented by the transpose convolution with stride = 2.

Model 2 is a "U"-structured model using residual connections and adding a channel attention (CA) block to the skip connection. Down-sampling strategy is the max-pooling, and up-sampling strategy is the transpose convolution with stride = 2.

Model 3 is a "U"-structured model using residual connections and replacing the bottom module of the "U"-shape network with a hybrid dilated attention convolutional (HDAC) layer. Down-sampling and up-sampling are same as Model 2.

The initial number of channels was set to 64 for LiTS 2017 and 32 for Lung dataset, DBS2018, ISBI 2012. Tables 6–9 show the performance comparison among Model 1, Model 2, Model 3 and our proposed model on the above four datasets, respectively.

According to Tables 6–9, we can see inserting the channel attention block in the skip connection and replacing the bottom module attained finer results than the standard U-Net. The experiments show that our proposed network has a more accurate and precise segmentation output than the ordinary U-Net. After the skip connection, there are two features to be combined. One is from the former decoder layer and the other

is from the matching encoder layer. In the ordinary U-Net, these two features are directly combined with a concatenation function. In this model, we employ the channel attention (CA) block to combine encoding and decoding features. After the channel compression, the encoded feature contains more local information from the input sample and global information of the channel. The improved hybrid dilated attention convolutional (HDAC) layer also assists the decoded feature to possess more semantic information of the input sample. The interaction between encoding and decoding features may form a group of feature maps with ample global information and semantic information. In addition, we add residual short connections to speed up the network learning process and facilitate network optimization. To evaluate the effect of the residual short connections, we trained the network with and without residual short connections separately on the DSB 2018 dataset. Fig. 15(a) displays the training and validation dice coefficients of the proposed HDA-ResUNet on the DSB 2018 dataset without residual short connections, and Fig. 15(b) displays the identical evaluation metric of the network with residual short connections. The HDA-ResUNet becomes stable after 90 epochs without residual short connections, while the network with residual short connections becomes stabilized after about 20 epochs. That is, residual short connections speed up the convergence of the network by a factor of 4.5.

4. Discussion

We design a residual convolutional neural network with a self-attention mechanism and a hybrid dilated convolution based on the U-Net to achieve the medical segmentation accurately. We evaluate networks on CT images (LiTS 2017, Lung dataset) and microscopic images (DSB 2018, ISBI 2012). The proposed network gains better results compared to the U-Net. The dice global improved by 0.85% for liver and 2.57% for tumor respectively on LiTS 2017. The dice improved by 2.66% on the Lung dataset and 0.69% on DSB 2018. The info score improved by 1.42% on ISBI 2012. Although the proposed method improves performance, the lack of medical data related to microscopic images limits the segmentation effect. And the liver dice global of 94.9 is slightly inferior to 96.3 in Ref. [33], because [33] employed two steps to complete liver segmentation. Firstly, they used a 2D network to roughly segment the liver and locate the approximate location of the liver, and then fine segmentation of liver was performed by a 3D network. Despite achieving better results, it also required more training time. The dice global and dice per case of tumor segmentation were 0.795, 0.595 in Ref. [33]. Ours were 0.799, 0.653. The former sent $128 \times 128 \times 32$ patches into 3D network to segment tumor, which leaded to insufficient intra-slice information to some extent, resulting in poor segmentation. HDA-ResUNet employed a 2D network, which lost some inter-slice information. However, the designed CA block and HDAC layer helped the proposed network to exploit the intra-slice information to a certain extent, thus the network obtained better tumor segmentation results. LiTS competition aims to accurately segment tumors to assist physicians' diagnosis, and thus our study is meaningful.

In the future, we expect to work on 3D networks to incorporate inter-slice information while seeking a balance between spatial dimensions and computational resources. In addition, we consider finding a more efficient method of data reading to reduce training time.

5. Conclusion

In this work, we propose the HDA-ResUNet for medical image segmentation. In order to obtain multi-scale information and global information to capture more favorable information and obtain more accurate segmentation results, we added channel attention block in the skip connection, and superseded the bottom layer with a hybrid dilated attention convolutional layer. Experimental results on four open benchmark datasets show that our model has significantly fewer parameters and is well segmented compared with U-Net.

Declaration of competing interest

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of the manuscript entitled "Hybrid Dilation and Attention Residual U-Net for Medical Image Segmentation".

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 61862044 and 51765042. This work was also supported by Jiangxi Natural Science Foundation under grant 20192BAB207015 and 20171ACB20007. This work was also supported by the Innovation Fund Designated for Graduate Students of Jiangxi Province under grant YC2020–S091.

References

[1] Y. Xie, Z. Zhang, M. Sapkota, L. Yang, Spatial clockwork recurrent neural network for muscle perimysium segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 185–193.
[2] M. Li, C. Wang, H. Zhang, G. Yang, Mv-ran, Multiview recurrent aggregation network for echocardiographic sequences segmentation and full cardiac cycle analysis, Comput. Biol. Med. 120 (2020), 103728.
[3] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[4] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[5] 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 (4) (2017) 834–848.
[6] H. Noh, S. Hong, B. Han, Learning deconvolution network for semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1520–1528.
[7] W. Sun, R. Wang, Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm, Geosci. Rem. Sens. Lett. IEEE 15 (3) (2018) 474–478.
[8] A.G. Roy, S. Conjeti, N. Navab, C. Wachinger, A.D.N. Initiative, et al., Quicknat: a fully convolutional network for quick and accurate segmentation of neuroanatomy, Neuroimage 186 (2019) 713–727.
[9] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, D. Feng, Dermoscopic image segmentation via multistage fully convolutional networks, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 64 (9) (2017) 2065–2074.
[10] 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.
[11] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[12] 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.
[13] 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.
[14] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, Bi-directional convlstm u-net with densley connected convolutions, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, 0–0.
[15] D. Li, D.A. Dharmawan, B.P. Ng, S. Rahardja, Residual u-net for retinal vessel segmentation, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1425–1429.
[16] A. Kermi, I. Mahmoudi, M.T. Khadir, Deep convolutional neural networks using u-net for automatic brain tumor segmentation in multi-modal mri volumes, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 37–48.
[17] S.K. Devalla, P.K. Renukanand, B.K. Sreedhar, G. Subramanian, L. Zhang, S. Perera, J.-M. Mari, K.S. Chin, T.A. Tun, N.G. Strouthidis, et al., Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images, Biomed. Opt Express 9 (7) (2018) 3244–3265.
[18] J. Zhang, Z. Jiang, J. Dong, Y. Hou, B. Liu, Attention gate resu-net for automatic mri brain tumor segmentation, IEEE Access 8 (2020) 58533–58545.
[19] B.J. Bhatkalkar, D.R. Reddy, S. Prabhu, S.V. Bhandary, Improving the performance of convolutional neural network for the segmentation of optic disc in fundus images using attention gates and conditional random fields, IEEE Access 8 (2020) 29299–29310.

[20] 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.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[22] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[23] 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.
[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[25] S. Ioffe, C. Szegedy, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015 arXiv preprint arXiv:1502.03167.
[26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[27] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, G. Cottrell, Understanding convolution for semantic segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1451–1460.
[28] D.P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[29] I. Arganda-Carreras, S.C. Turaga, D.R. Berger, D. Cireşan, A. Giusti, L. M. Gambardella, J. Schmidhuber, D. Laptev, S. Dwivedi, J.M. Buhmann, et al., Crowdsourcing the creation of image segmentation algorithms for connectomics, Front. Neuroanat. 9 (2015) 142.
[30] LiTS. https://competitions.codalab.org/competitions/15595, 2017.
[31] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics Gems IV, 1994, pp. 474–485.
[32] N.M. Aboelenein, P. Songhao, A. Koubaa, A. Noor, A. Afifi, Httu-net: hybrid two track u-net for automatic brain tumor segmentation, IEEE Access 8 (2020) 101406–101415.
[33] Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, Ra-unet, A hybrid deep attention-aware network to extract liver and tumor in ct scans, Front. Bioeng. Biotechnol. 8 (2020) 1471.
[34] https://www.kaggle.com/kmader/finding-lungs-in-ct-data.
[35] DSB. https://www.kaggle.com/c/data-science-bowl-2018, 2018.
[36] 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, Springer, 2018, pp. 3–11.
[37] ISBI. http://brainiac2.mit.edu/isbi_challenge/, 2012.