Neurocomputing 486 (2022) 135–146

Contents lists available at ScienceDirect

Neurocomputing

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

Multimodal medical image segmentation using multi-scale context-aware network

Xue Wang [a,b], Zhanshan Li [a,b], Yongping Huang [a,b,*], Yingying Jiao [a,b,*]

[a] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[b] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University 130012 Changchun, China

ARTICLE INFO

Article history:
Received 16 July 2021
Revised 5 October 2021
Accepted 7 November 2021
Available online 15 November 2021

Keywords:
Medical image segmentation
Atrous convolution
Dense skip connection
Multi-scale context fusion
Multimodality

ABSTRACT

Multimodal medical image segmentation with different imaging devices is a key but challenging task in medical image visual analysis and reasoning. Recently, U-Net based networks achieved considerable success in semantic segmentation of medical image. However, U-Net utilizes a skip-connection to connect two symmetric encoder and decoder layers. Although the single granularity information of the encoder layer is preserved through skip connection, the rich multi-scale spatial information is ignored, which greatly affects its performance in the segmentation task. In this paper, a multi-scale context-aware network (CA-Net) for multimodal medical image segmentation is proposed, which captures rich context information with dense skip connection and assigns distinct weights to different channels. CA-Net consists of four key components, namely encoder module, multi-scale context fusion (MCF) module, decoder module, and dense skip connection module. The proposed MCF module extracts multi-scale spatial information through a spatial context fusion (SCF) block, and learn to balance channel-wise features through a Squeeze-and-Excitation (SE) block. Extensive experiments demonstrate that our model achieves state-of-the-art performance on three benchmark datasets of different modalities, including skin lesion segmentation in dermoscopy, lung segmentation in CT images, and blood vessel segmentation in retina images.

1. Introduction

Image segmentation is a challenging task in visual media reasoning. Due to the development of medical imaging equipment, intelligent visual computing over multi-modal data to assist clinical diagnosis has attracted public attention in medical field. Multimodal medical image segmentation lays the ground for medical image analysis and is useful in a wide range of applications. For example, vessels segmentation [2,23,47] in retinal images assists ophthalmologists in evaluating eye diseases, and brain tumor segmentation [20,40,44] in magnetic resonance imaging (MRI) lays the foundation for early diagnosis.Accurate medical image segmentation delivers doctors with precise interpretation of medical images, helping doctors make more accurate judgement and develop effective treatment plans [2,3,14,46,51].

Pioneering works for medical image segmentation rough-ly cast into four categories, namely superpixel-based graph cut methods [43], level set-based methods [55,56], clustering-based methods [25,31], and region growing-based methods [22,34]. These conventional methods typically utilize human expert knowledge to manually design features for segmentation. Unfortunately, it is inherently difficult to design effective features that are suitable for a dataset may be inefficacy for another, especially in multimodal data types. Therefore, a general method that can automatically extract features for segmentation and is extendable to datasets of different modalities is much coveted.

Recently, with the development of convolutional neural network (CNN), it achieves considerable success in object detection [11,32], human pose estimation [26], cross-media retrieval [45,48,49], and semantic segmentation [53]. It also results in the rediscovery of GNNs, which have excellent performances on many deep learning tasks [27,54]. Deep learning-based segmentation methods overcome the limitation of hand-crafted features and show their great potential in learning features automatically. Fully convolutional neural network (FCN) [28] greatly promotes the development of segmentation models. More network architectures such as U-Net [33], PSPNet [53], SegNet [4], and Deeplab versions [6–8] are then proposed. Among them, U-Net achieves state-of-the-art performance in medical image segmentation, which uses a skip-connection encoder-decoder structure to extract features. Thereafter, various extensions of U-Net are presented for medical image segmentation tasks [21,24,30,35,57].

* Corresponding authors at: College of Computer Science and Technology, Jilin University, Changchun 130012, China.
E-mail addresses: hyp@jlu.edu.cn (Y. Huang), jiaoying2013@126.com (Y. Jiao).

https://doi.org/10.1016/j.neucom.2021.11.017
0925-2312/

The challenges of medical image segmentation lie in its multi-modality and the irregular shapes of different target objects in different datasets. Moreover, target objects usually have variant scales and occupy a small part of the whole image as shown in Fig. 1. For instance, as demonstrated in the left of Fig. 1, the skin lesion possesses an irregular shape and its scale can greatly vary in dermoscopy images. Similarly, as illustrated in the middle of Fig. 1, the intensity of surrounding tissues is inhomogeneous and the boundary is complicated in lung CT images. In the right of Fig. 1, the vessels in retinal image are tiny and their shapes are irregular. These situations are common in other medical image segmentation tasks. Therefore, context information from different scales is of great importance for analyzing objects in medical images, especially in challenging scenarios.

State-of-the-art medical image segmentation frameworks are mostly based on U-Net or its variants, which utilize skip-connection encoder-decoder architectures. Although U-Net and its extensions have achieved dramatic performance in medical image segmentation, they still suffer from the following limitations in preserving context information. First, the encoder of U-Net stacks a set of convolutional operations in a cascade way to gradually learn higher-level feature representations. The conventional convolution operation only extracts local information, which is not beneficial for the medical image segmentation that usually depends on a wide range of context information. Meanwhile, a set of pooling operations are adopted to reduce the spatial sizes of feature maps layer by layer. The consecutive pooling operations cause the loss of more spatial information during encoding, which is not conducive to the restoration of features in the decoding process. Second, when restoring the spatial information of feature maps in the decoder, a skip connection only connects a pair of symmetric encoder and decoder layers, which neglects important information like object location and boundaries in low-level semantic feature maps.

Motivated by this, we propose CA-Net (multi-scale Context Aware Network) for multimodal medical image segmentation. The CA-Net fuses rich context information including multi-scale spatial contexts and different channel-wise contexts. To evaluate the effectiveness of the network, CA-Net is studied on three benchmark datasets of different modalities. Extensive experiments demonstrate that our method achieves state-of-the-art performance with fewer parameters on all these datasets.

The main contributions of this paper can be summarized as follows:

- A novel multi-scale context fusion (MCF) module is proposed, consisting of a spatial context fusion (SCF) block and a Squeeze-and-Excitation (SE) block, which captures more spatial semantic features from multiple scales and automatically assigns distinct weights to different channel-wise context information.
- Dense skip connections are proposed to preserve multiple levels of detailed information from pre-feature maps, and are integrated with the encoder-decoder str-ucture.
- We evaluate the proposed multi-scale context-aware model on three benchmark medical image segmentation datasets of different modalities, including skin lesion segmentation in dermoscopy, lung segmentation in CT images, and blood vessel segmentation in retina images. Empirically, our method achieves the new state-of-the-art performance.

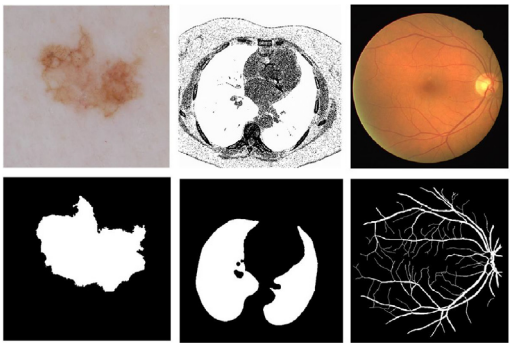


Fig. 1. Different medical image segmentation tasks: the first column is skin lesion segmentation in dermoscopy,the second column is lung segmentation in CT images, and the third column is blood vessel segmentation in retina images.

The paper is organized as follows: Section 2 reviews the related work. Section 3 introduces our method in detail. Section 4 presents the experimental results and findings, while Section 5 summarizes the paper and discusses future directions.

2. Related work

Multimodal semantic segmentation in different medical image analysis is an active topic and is full of challenges. Recently, deep learning-based approaches have achieved excellent performance in medical image segmentation. Technically, we classify them into three categories as below.

2.1. Encoder-Decoder architectures

Using encoder-decoder architectures is one of the most popular strategies in semantic segmentation networks. U-Net, proposed by Ronneberger et al. [33], utilized a skip-connection encoder-decoder architecture to extract semantic features through continuous convolution and pooling operations. It works well even with simple network architecture in different medical segmentation tasks. Md Zahangir Alom et al. [2] proposed R2U-Net by building upon the power of U-Net, Residual Network, and RCNN, obtaining better performance with the same amount of computation as U-Net for different medical image segmentation tasks. Nabil Ibtehaz et al. [21] revised the U-Net architecture. They proposed MultiRes blocks to extract semantic information from multiple scales. They also utilized Res paths to alleviate semantic gap between two symmetric encoder and decoder layers. For better feature propagation and reuse, BCDU-Net [3] was proposed by combining the advantages of U-Net, bi-directional ConvLSTM, and the dense convolutions.To capture small and complex variations in medical images, PyDiNet [13] was proposed by introducing a pyramid dilated module (PDM). In [12], a recurrent framework of two interconnected networks by using the context feedback loop was proposed for robust medical image segmentation. Additionally, to process 3D MRI volumes, Milletari et al. [29] proposed V-Net on the basis of U-Net, an end-to-end 3D medical image segmentation network.

The common issue of these encoder-decoder based architectures is that, they usually extract semantic features progressively layer by layer in the encoding procedure, and recover the size and detailed information of the feature map step by step in the decoding process, which achieves end-to-end pixels-wise segmentation but is insufficient in using rich context information. In medical image segmentation, we usually need to consider more context information around the area to be segmented.

2.2. Attention mechanism

Recently, attention-based models have achieved considerable performance in semantic segmentation [10,19]. These algorithms apply the attention modules to the feature maps obtained from

convolution neural networks. To highlight the salient features, Ozan Oktay et al. [30] adopted a novel attention gate (AG) in the U-Net architecture. Squeeze-and-Excitation network, proposed by Jie Hu et al. [18], utilized a lightweight gating mechanism named SE block to model channel-wise relationships and enhance the representational power of the network. Yucheng Shu et al. [36] introduced AFT-Net, which applied a multi-stream encoder to enhance the feature representation ability, and an attentional fusion algorithm to build visual correlation between different layers. Jieneng Chen et al. [5] proposed TransUNet by leveraging the advantages of transformers that with self-attention mechanisms to boost the performance on medical image segmentation.In [37], a novel bi-directional seed attention module (BSA) was proposed for interactive segmentation,which highlighted the context information of the feature by alternately updating the seed map and the feature map.

Attention mechanism suppresses unrelated regions, and enhances the salient regions by reweighting the feature maps. In this way, the segmentation performance of the network is improved. Common attention mechanisms usually focus on the spatial and channel-wise relationships, which only capture local dependencies, failing to make the best of multi-scale context information around the target region. Therefore, multi-scale self-attention is crucial to capture richer feature dependencies.

2.3. Context-based methods

Modeling context information is one of the important components in semantic segmentation. Obtaining context information for deep convolution neural networks is challenging in requiring to consider both local and global dependencies. Liang-Chieh Chen et al. [7] proposed atrous spatial pyramid pooling (ASPP) to robustly segment objects by capturing image context at multiple scales. Zaiwang Gu et al. [14] proposed a context extractor module and integrated it with encoder-decoder structure for medical image segmentation. In [16], non-local context encoder module was introduced to model global spatial dependencies and channel-wise feature map attention. It could be embedded into other network structures, and achieved high robustness and excellent segmentation performance in biomedical image segmentation. For robust segmentation on medical images, a Crossover-Net [50] was proposed by taking the crossover-patch as complement contextual information.To boost the 3D medical image segmentation accuracy, Zhang et al. [52] proposed a 3D context residual network by utilizing the context residual module to perceive inter-slice context.

In this paper, a multi-scale context-aware network (CA-Net) is proposed, which captures rich context information with multi-scale convolutions, dense skip connections, and attention on channel-wise features.

3. Method

The proposed multi-scale context-aware network, CA-Net, is extended from U-Net and consists of four components, namely encoder module, multi-scale context fusion (MCF) module, decoder module, and dense skip connection module. The overall architecture of CA-Net is demonstrated in Fig. 2. Specifically, the image first goes through a U-Net encoder that consists of several convolution and pooling layers. Then, the features are fed into a novel MCF module, which fuses contexts from multiple scales using atrous convolution and reweights information from different channels using a squeeze and excitation block. Finally, the feature maps are passed to a U-Net decoder to yield the segmentation results. Instead of using plain skip connections, dense skip connections are engaged to enrich the context of a decoder layer with information from the symmetric encoder layer and all lower-level encoder layers. In what follows, we will introduce the proposed MCF (multi-scale context fusion) module and dense skip connections, respectively.

3.1. Multi-scale context fusion module

Our proposed MCF module consists of the spatial context fusion (SCF) block, Squeeze-and-Excitation (SE) block and a residual connection. As shown in Fig. 3, MCF absorbs features of the last layer of the U-Net encoder. The SCF block fuses contexts of different scales utilizing atrous convolution with multiple atrous rates. The contexts are then fed into the Squeeze-and-Excitation block to assign distinct weights to the 512 different channels.

The idea of multi-scale context fusion has advantages in medical image feature learning. First, the SCF block consists of multiple branches, and each branch utilizes convolutional operations of different kernel sizes to extract the points of interest from different scales, and it enables the network to learn more context information by fusing spatial information of different scales. Moreover, in SE block, weight learning across channels strengthens the power of contextual semantic representation in high-level semantic feature maps.

3.1.1. SCF block

To address the limitation of using fixed-scale convolutional operations in segmenting objects of interest, we design the SCF block, which also improves robustness of the network by analyzing from different scales. In this block, we utilize convolutional operations of different kernel sizes in multiple branches to extract features of different scales, which enables the network to learn more context information by fusing spatial information of different granularities.

Inspired by Inception [41,42] and ASPP [6,38], the structure of SCF block is shown in Fig. 3(a). There are four branches in SCF block, where each branch utilizes convolutional operations of a different kernel size. To reduce the amount of model calculation while enlarging the field of filter views, we adopt atrous convolution [17]. The rates of atrous convolution in SCF block are 1,3, and 5, and the receptive fields of the branches are 3, 7, 11 and 19 respectively. By employing multiple paralleled filters with different receptive fields in the SCF block, we can obtain context information from multiple scales for each pixel in the feature map.

3.1.2. SE block

After fusing spatial information from multiple scales in SCF block, to strengthen the power of contextual semantic representation in the high dimensional feature maps, SE block [18] is introduced, which is shown in Fig. 3(b). We employ SE block to focus attention on channel dependencies by explicitly modelling interdependencies between channels. The SE block includes squeeze and excitation units. In sque-eze unit, global average pooling is utilized to build channel-wise statistics. More specifically, let $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ be the input feature maps, and $\mathbf{z} \in \mathbb{R}^{1\times 1\times C}$ be the statistics from global average pooling. Denoting $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_C]$ as the set of input feature maps, where $\mathbf{x}_c \in \mathbb{R}^{H\times W}$. $\mathbf{z} = [z_1, z_2, \ldots, z_C]$ denotes the learned set of statistics, where z_c is calculated by:

$$z_c = F_{sq}(\mathbf{x}_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c(i,j), \tag{1}$$

where $F_{sq}(\mathbf{x}_c)$ stands for global average pooling operation. After global average pooling operation, the dimensions of $\mathbf{X}$ change from $H \times W \times C$ to $1 \times 1 \times C$. The output of the squeeze unit is a global description of $\mathbf{X}$.

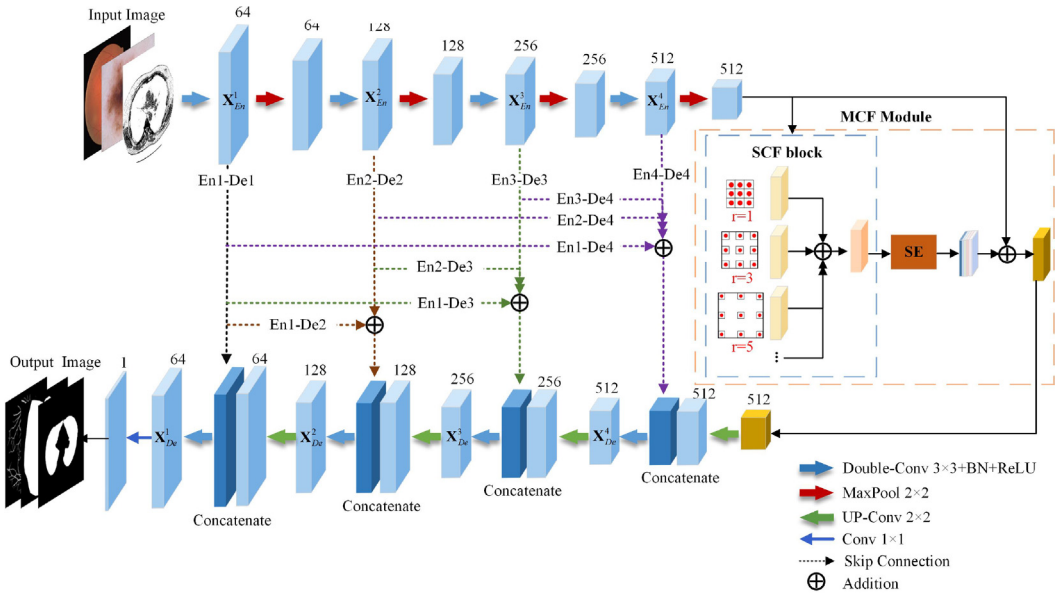


Fig. 2. CA-Net with MCF module and dense skip connections..

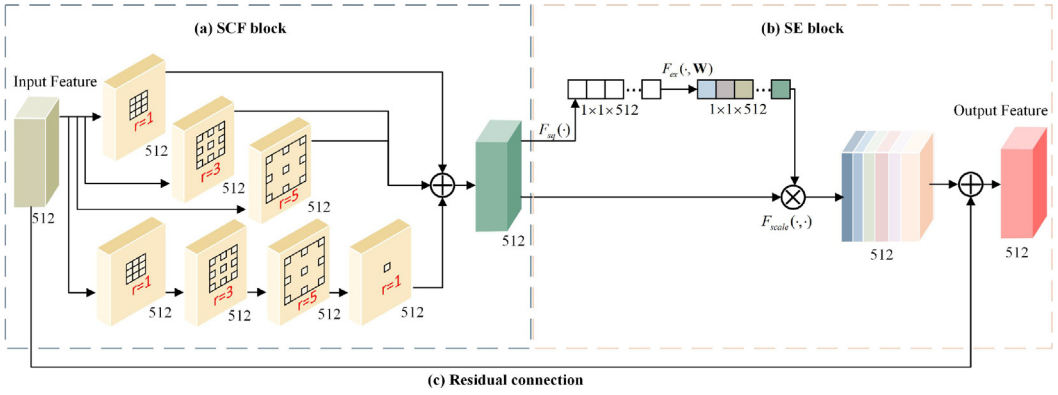


Fig. 3. An illustration of the MCF module. It consists of a SCF block, a SE block, and a residual connection..

To capture channel-wise dependencies, the excitation unit is defined as follows:

$$\mathbf{s} = F_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})), \tag{2}$$

where $F_{ex}(\mathbf{z}, \mathbf{W})$ is excitation operation, σ denotes sigmoid function, δ represents ReLU function, $g(\mathbf{z}, \mathbf{W})$ denotes a function which contains two FC operations with $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}, \mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$. r is reduction ratio that mainly used to reduce the number of network parameters, here we set it to 16 [18]. The output of the unit is obtained by rescaling $\mathbf{X}$ with the activations $\mathbf{s}$, which is given by:

$$\tilde{\mathbf{x}}_c = F_{scale}(\mathbf{x}_c, s_c) = s_c\mathbf{x}_c, \tag{3}$$

where $F_{scale}(\mathbf{x}_c, s_c)$ denotes channel-wise multiplication between the scalar s_c and the feature map $\mathbf{x}_c \in \mathbb{R}^{H\times W}$, and $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_C]$.

Through squeeze and excitation units, the SE block introduces self-attention mechanism on channels of the feature maps, which can dynamically learn weight information of the channels. Therefore, the power of contextual semantic representation in high dimensional feature maps is enhanced by employing SE block, which is not confined to the local spatial information that the SCF block is responsible to.

Finally, inspired by ResNet [15], we add a shortcut residual connection, as shown in Fig. 3(c), to avoid the gradient vanishing and accelerate the network convergence. At the same time, adding a residual connection can also avoid the loss of spatial context information and preserve more details of pre-feature maps.

3.2. Dense skip connections

The skip connections between symmetric encoder and decoder layers in the U-Net architecture is an ingenious idea, which preserves the spatial information that gets lost during the encoding process. However, some details like position and boundary information are still lost since a decoder layer is only connected to a single encoder layer.

To solve this problem, and preserve more detailed information from pre-feature maps in the encoder, dense skip connections between the encoder and decoder are proposed. Specifically, each decoder layer in CA-Net combines details from both the symmetric encoder layer and all upper-level encoder layers, preserving more spatial context information and facilitating the reuse of multi-level features in restoring image.

We denote the feature maps from the i-th encoder layer as $\mathbf{X}_{En}^{i}$, and those from the i-th decoder layer as $\mathbf{X}_{De}^{i}$. Take $\mathbf{X}_{De}^{3}$ as an example, the decoder feature maps $\mathbf{X}_{De}^{3}$ fuse lower-level detailed information from encoder feature maps $\mathbf{X}_{En}^{1}$ and $\mathbf{X}_{En}^{2}$, and high-level semantic information from symmetric encoder feature maps $\mathbf{X}_{En}^{3}$

by pixel-wise addition. Fig. 4 illustrates how to construct the feature maps of $\mathbf{X}_{De}^{3}$. In order to realize the pixel-wise addition of different encoder layer feature maps, we need to unify the size and number of channels of different feature maps by applying max pooling and convolution operations. Specifically, we perform 4×4 max pooling with stride 4 on $\mathbf{X}_{En}^{1}$ to unify the size of $\mathbf{X}_{En}^{1}$ and $\mathbf{X}_{En}^{3}$, and use 1×1 convolution operation on $\mathbf{X}_{En}^{1}$ to unify the number of channels for $\mathbf{X}_{En}^{1}$ and $\mathbf{X}_{En}^{3}$. In a similar way, we perform 2×2 max pooling with stride 2 on $\mathbf{X}_{En}^{2}$ to unify the size of $\mathbf{X}_{En}^{2}$ and $\mathbf{X}_{En}^{3}$, and utilize 1×1 convolution operation on $\mathbf{X}_{En}^{2}$ to unify the number of channels for $\mathbf{X}_{En}^{2}$ and $\mathbf{X}_{En}^{3}$. At last, we fuse spatial information of $\mathbf{X}_{En}^{1}, \mathbf{X}_{En}^{2}$ and $\mathbf{X}_{En}^{3}$ by pixel-wise addition operation. In this way, we get feature maps containing different levels of semantic information, then we concatenate them with feature maps obtained by up-sampling decoder layer $\mathbf{X}_{De}^{4}$. Further, we perform two consecutive 3×3 convolution operations, each followed by a batch normalization and ReLU activation function.The dense skip connections can be formulated as follows:

$$\mathbf{X}_{De}^{i} = \begin{cases} \mathrm{C}_1(\mathbf{X}_{En}^{i} \bullet \mathrm{U}(\mathbf{X}_{De}^{i+1})), i = 1 \\ \mathrm{C}_1((\sum_{k=1}^{i-1} \mathrm{C}_2(\mathrm{P}(\mathbf{X}_{En}^{k})) + \mathbf{X}_{En}^{i}) \bullet \mathrm{U}(\mathbf{X}_{De}^{i+1})), 1 < i < 4 \\ \mathrm{C}_1((\sum_{k=1}^{i-1} \mathrm{C}_2(\mathrm{P}(\mathbf{X}_{En}^{k})) + \mathbf{X}_{En}^{i}) \bullet \mathrm{U}(\mathbf{X}_{mcf})), i = 4 \end{cases} \tag{4}$$

where function $C_1(.)$ represents two consecutive 3×3 convolution operations, and $C_2(.)$ represents a 1×1 convolution operation, each convolution operation followed by a batch normalization and ReLU activation function. P(.) stands for max pooling operation with kernel size 2^{i-k}, U(.) denotes up-sampling operation with a sampling factor of 2, and •is concatenation. $\mathbf{X}_{mcf}$ represents the feature maps obtained by the MCF module. The architectural and operational details of each path in dense skip connections are demonstrated in Table 1.

Table 1
Dense Skip Connections (DSC) Architecture Details.

DSC	path	Operations
DSC1	En1-De1	-
DSC2	En1-De2	MaxPool2 × 2 128,Conv1 × 1
	En2-De2	-
DSC3	En1-De3	MaxPool4 × 4 256,Conv1 × 1
	En2-De3	MaxPool2 × 2 256,Conv1 × 1
	En3-De3	-
DSC4	En1-De4	MaxPool8 × 8 512,Conv1 × 1
	En2-De4	MaxPool4 × 4 512,Conv1 × 1
	En3-De4	MaxPool2 × 2 512,Conv1 × 1
	En4-De4	-

4. Experiments

4.1. Datasets and metrics

We engage three public benchmark datasets from medical domain to evaluate the proposed method CA-Net. Detailed descriptions of the datasets are summarized in Table 2.

To evaluate the performance of CA-Net, we adopt several evaluation metrics commonly used in image segmentation, including accuracy (*Acc*), sensitivity (*Sen*), specificity (*Spec*), and $F1-Scroe$. Mathematically, they are:

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{5}$$

$$Sen = \frac{TP}{TP + FN} \tag{6}$$

$$Spec = \frac{TN}{TN + FP} \tag{7}$$

$$Prec = \frac{TP}{TP + FP} \tag{8}$$

$$F1 - Score = \frac{2Prec * Sen}{Prec + Sen} \tag{9}$$

where *TP*, *TN*, *FP* and *FN* denote true positive, true negative, false positive and false negative, respectively. $F1-Score$ represents the harmonic mean of the precision and sensitivity.

4.2. Experimental setup

Our proposed CA-Net is implemented by using the public PyTorch platform. The training and testing are conducted on a

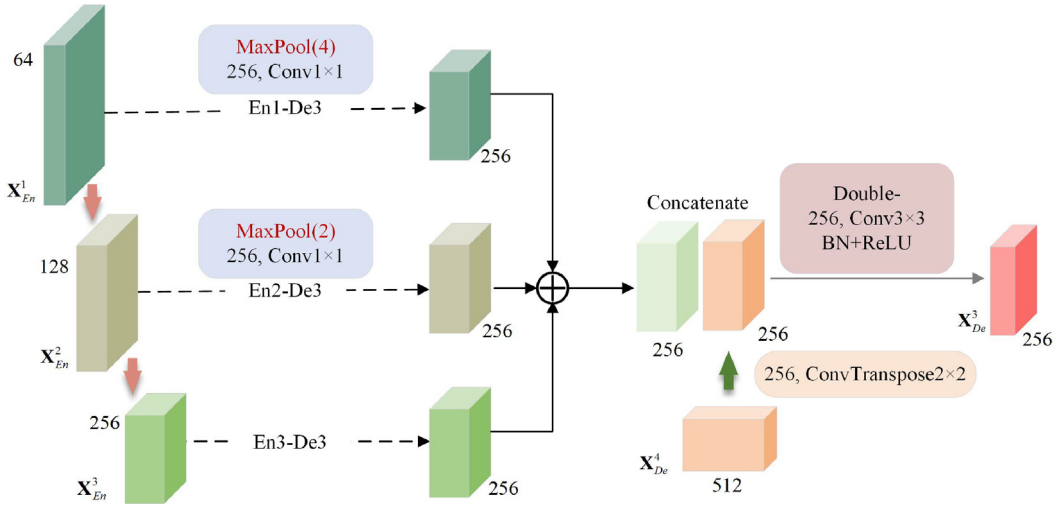


Fig. 4. An illustration of how to construct the feature maps of the third decoder layer $\mathbf{X}_{De}^{3}$.

Table 2
Overview of medical segmentation datasets used in our experiments.

Dataset	No. of Images	Modality	Original Resolution	Input Resolution
ISIC 2018 [9]	2594	Dermoscopy	variable	448 × 448
Lung [1]	267	CT images	512 × 512	512 × 512
DRIVE [39]	40	retina images	565 × 584	256 × 256

Table 3
Comparison results on ISIC 2018 dataset.

Methods	*F1 − Score*	*Acc*	*Sen*	*Spec*
U-Net [33]	0.647	0.890	0.708	0.964
Atten U-Net [30]	0.665	0.897	0.717	0.967
R2U-Net [2]	0.679	0.880	0.792	0.928
BCDU(d = 1) [3]	0.847	0.936	0.783	0.980
BCDU(d = 3) [3]	0.851	0.937	0.785	0.982
CA-Net	**0.868**	**0.957**	**0.855**	**0.985**

Ubuntu 16.04 LTS 64-bit system with the NVidia GeFo-rce GTX 1080 Ti discrete graphics cards, which has 64 Gigabyte memory. In ISIC 2018 dataset, we divide the dataset by following the setting of [2]. In lung segmentation dataset, we use 80% of the images for training and the remaining 20% for testing [14]. The DRIVE dataset contains 40 color retina images, consisting of 20 training samples and 20 testing samples. Considering the small number of training samples, here we perform data augmentation on training samples, including rotation with a maximum rotation angle of 10, horizontal flip, vertical flip, image zoomed by 0.9 times, image cropped by 0.8 times and image brightness adjustment randomly. In this way, we get 2,000 samples. We use 90% of them for training, 10% for validation. During training,binary cross-entropy and Adam are employed as the loss function and optimizer, respectively. The batch size of ISIC 2018, Lung, and DRIVE datasets is 6,4,6 respectively. The weight decay is $1e^{-4}$, and the initial learning rate is 0.0001. The maximum epoch is 200. We adopt the update strategy of learning

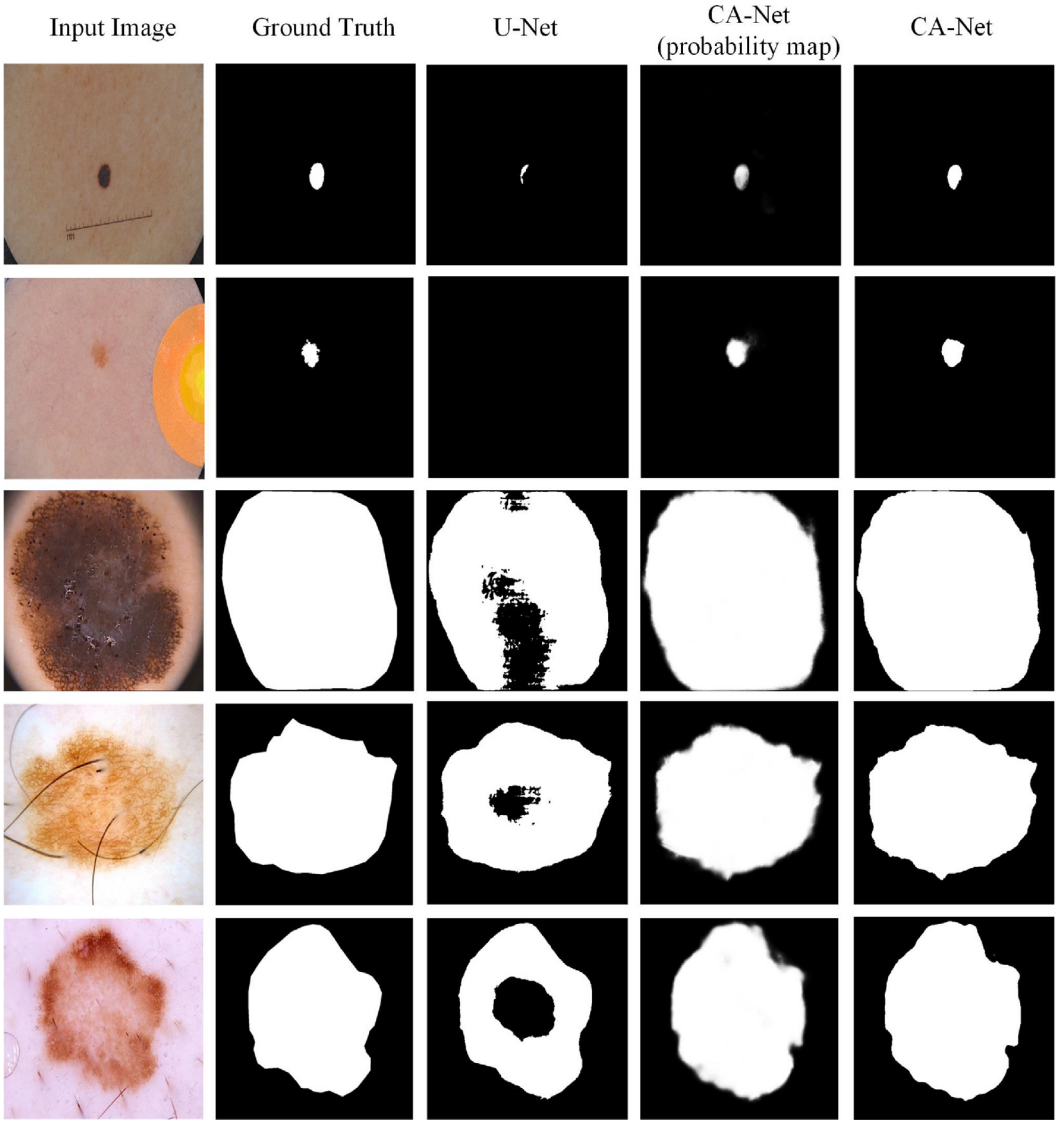


Fig. 5. Segmentation results of CA-Net on ISIC 2018..

rate,which shrinks by a factor of 0.95 when the train loss remains constant for 5 consecutive epochs. If the learning rate is less than $1e^{-6}$, we stop the training.

4.3. Comparison results

We present the experimental results and compare our method with state-of-the-art U-Net and its variants on three medical image segmentation datasets of different modalities.

4.3.1. Comparison on ISIC 2018 dataset

The first task is skin lesion segmentation. We perform the experiment on the ISIC 2018 dataset. Table 3 lists the quantitative results of different methods on ISIC 2018 dataset. As shown in the table, the proposed CA-Net achieves a substantial improvement on the four metrics. Fig. 5 presents some visual comparison results on ISIC 2018 dataset. From the figure, we can observe that the visual effect of CA-Net is superior to the U-Net in segmenting the areas of skin lesions, which are of irregular shapes, have variant scales, and are fuzzy at the border. Fig. 8(a) shows the ROC curves of CA-Net on ISIC 2018 dataset.

Table 4
Comparison results on Lung Segmentation dataset.

Methods	*F*1 − *Score*	*Acc*	*Sen*	*E*
U-Net [33]	0.875	0.939	0.974	0.139
Atten U-Net [30]	0.980	0.991	0.982	0.024
R2U-Net [2]	0.951	0.977	0.922	0.062
CE-Net [14]	-	0.990	0.980	0.038
CA-Net	**0.981**	**0.992**	**0.983**	**0.023**

4.3.2. Comparison on Lung Segmentation dataset

The second task is lung segmentation. We use 2D CT images for lung segmentation, which is provided by the Kaggle Data Science Bowl in 2017 [1]. Table 4 and Fig. 6 present comparison results on Lung segmentation dataset. The quantitative results in Table 4 show that the CA-Net achieves 0.981 in *F*1 − *Score*, 0.992 in *Acc*, 0.983 in *Sen*, and 0.023 in overlapping error denoted by *E* [14], which are better than the other methods listed in Table 4. Fig. 8 (b) shows the ROC curves of CA-Net on Lung Segmentation dataset.

4.3.3. Comparison on DRIVE dataset

Our experimental results on DRIVE dataset for blood vessel segmentation are shown in Table 5. Since most of the areas in blood

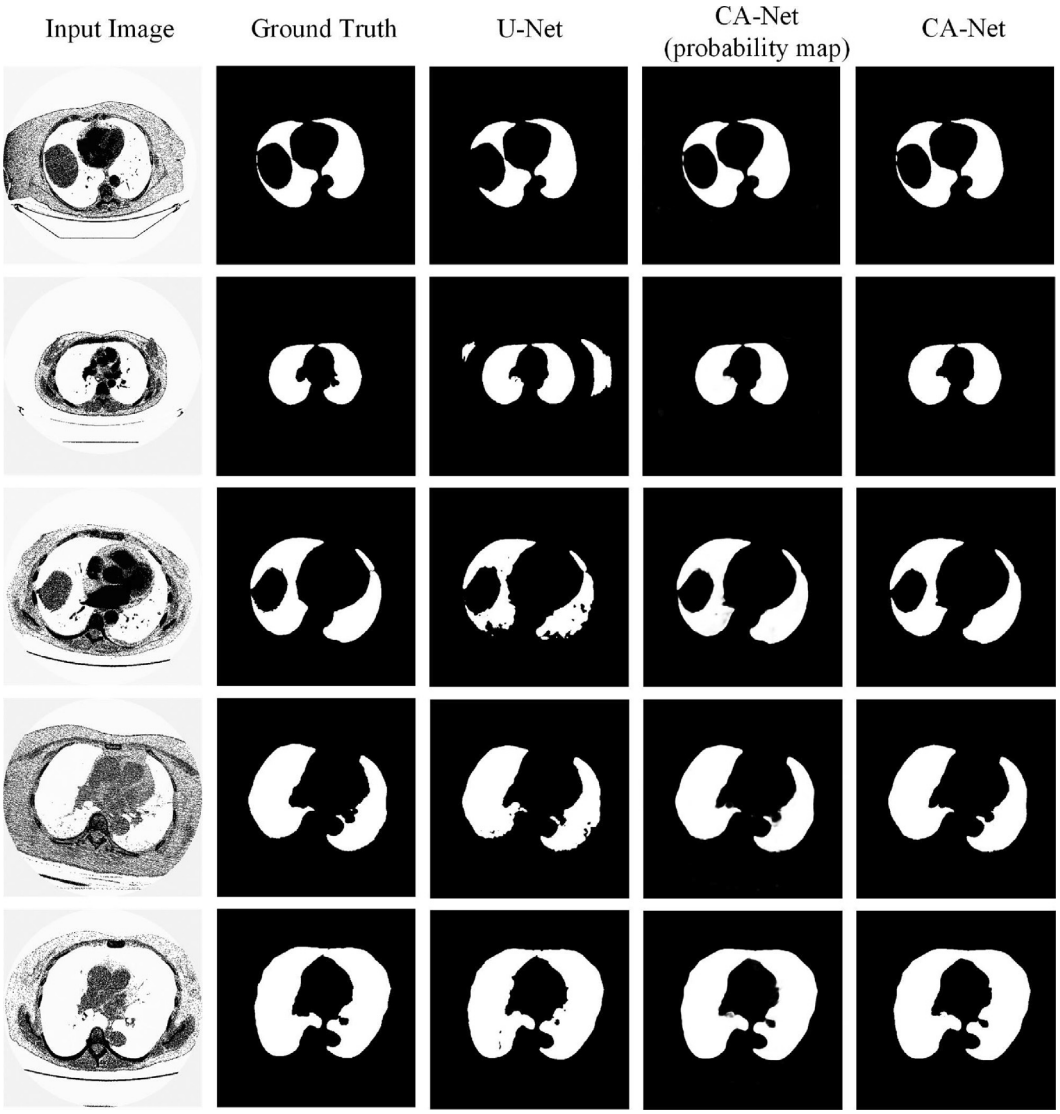


Fig. 6. Segmentation results of CA-Net on Lung Segmentation..

Table 5
Comparison results on DRIVE dataset.

Methods	$F1-Score$	Acc	Sen	$Spec$
U-Net [33]	0.8142	0.9531	0.7537	**0.9820**
Atten U-Net [30]	0.8181	0.9530	0.7910	0.9753
R2U-Net [2]	0.8171	0.9556	0.7792	0.9813
BCDU(d = 1) [3]	0.8222	0.9559	**0.8012**	0.9784
BCDU(d = 3) [3]	0.8224	0.9560	0.8007	0.9786
CA-Net	**0.8254**	**0.9561**	0.7934	0.9812

vessel image correspond to the background, and the blood vessel area only occupies a small fraction of the image, we take the segmentation threshold as 0.1. Compared with other methods, our proposed CA-Net achieves the $F1-Score$ of 0.8254 and the Acc of 0.9561, outperforming all other methods in Table 5. In order to investigate the superiority of CA-Net over the U-Net and its variants, no pre-processing and post-processing operations are performed, which causes the evaluation metrics Sen and $Spec$ are slightly lower than the BCDU-Net [3] and U-Net [33] by 0.78% and 0.08% respectively, while ensuring the comprehensive evaluation metrics of $F1-Score$ and Acc. Fig. 7 shows some examples for visual comparison on blood vessel segmentation. From the visual analysis, we observe that the proposed network is more precise than the U-Net in thin vessels segmentation. Fig. 8(c) shows the ROC curves of CA-Net on DRIVE dataset.

4.4. Ablation study

We analyze the efficacy of MCF module and dense skip connections in our CA-Net, by performing the ablation studies on ISIC 2018, Lung Segmentation, and DRIVE datasets. Table 6 lists the comparison results of $F1-Scroe, Acc, Sen$ and $Spec$ on the three datasets. Besides, to evaluate the efficiency of the proposed model, the number of model parameters and FLOPs are presented in Table 7.

(1) Ablation Study for MCF module

U-Net is the baseline of our CA-Net, we employ MCF module in the last convolutional layer of the encoder, aiming at extracting contextual semantic information from multiple spatial scales and multiple channels. Table 6 shows the segmentation results. After adopting the MCF module, CA-Net (**MCF + U-Net**) achieves a better segmentation performance. It can be seen that all the metrics are improved on three datasets, except that $Spec$ is slightly lower than U-Net on DRIVE dataset. However, the model parameters and

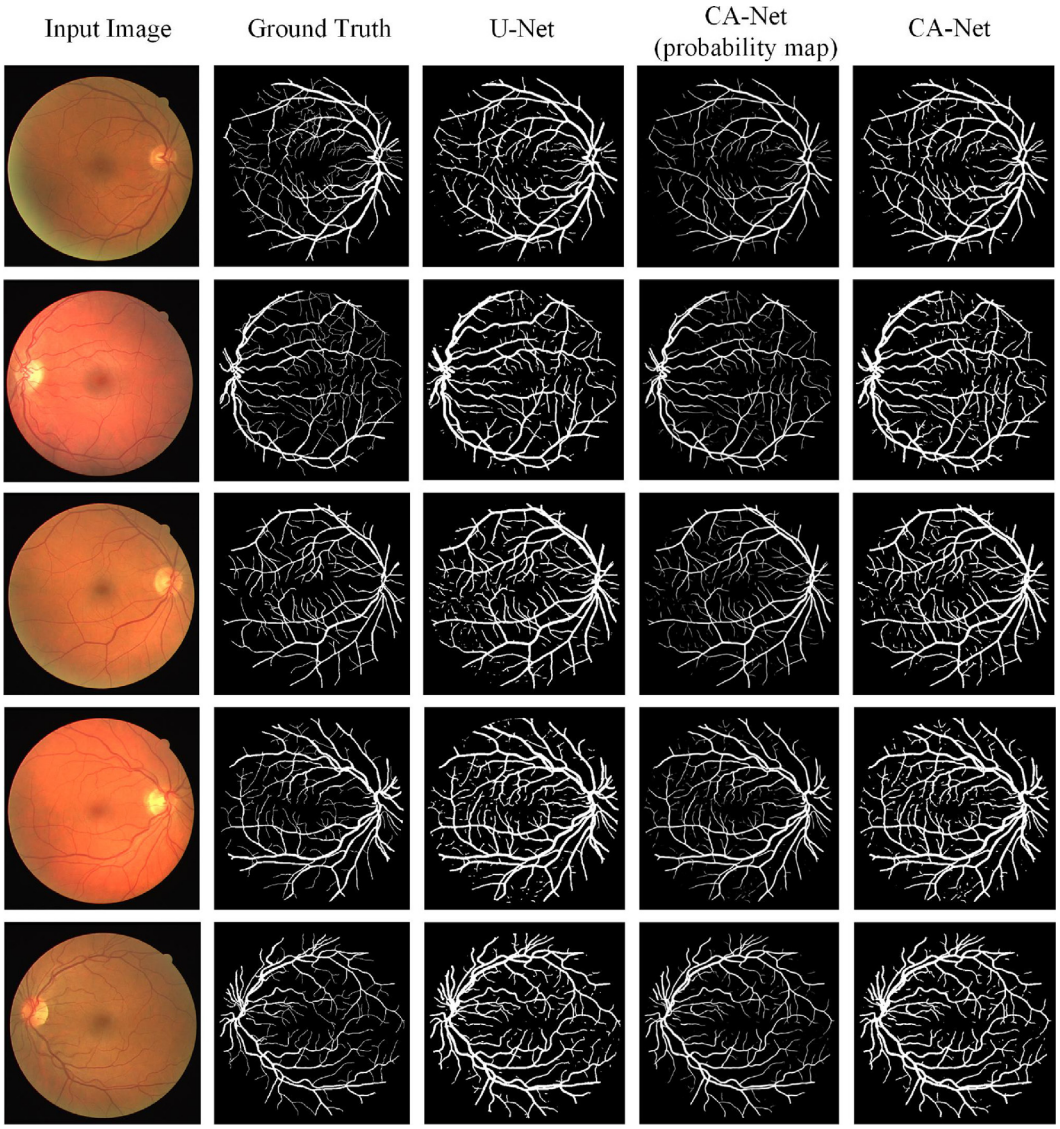
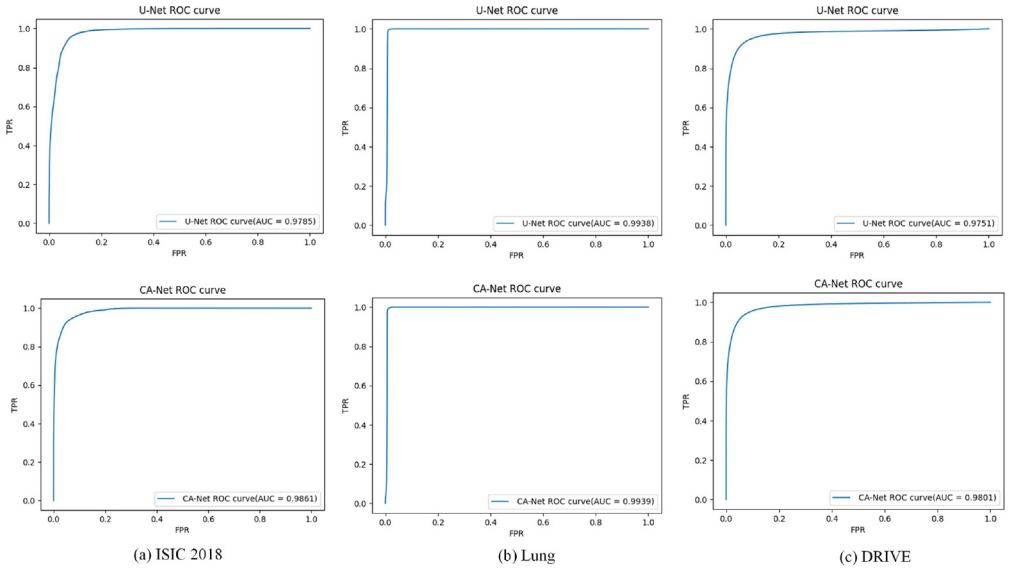


Fig. 7. Segmentation results of CA-Net on DRIVE..



(a) ISIC 2018 (b) Lung (c) DRIVE

Fig. 8. ROC curves of CA-Net on ISIC 2018, Lung and DRIVE datasets..

Table 6
Ablation study for each component on different datasets.

Methods	*F1-Score*	*Acc*	*Sen*	*Spec*
Dataset: ISIC 2018				
U-Net [33]	0.647	0.890	0.708	0.964
MCF + U-Net	0.853	0.950	**0.875**	0.971
MCF + atrous + U-Net	0.865	0.954	0.874	0.979
DSC + U-Net	0.830	0.950	0.829	0.970
MCF + atrous + DSC + U-Net (Ours)	**0.868**	**0.957**	0.855	**0.985**
Dataset: Lung Segmentation				
U-Net [33]	0.875	0.939	0.974	0.930
MCF + U-Net	0.981	0.991	0.982	0.994
MCF + atrous + U-Net	0.967	0.985	0.979	0.987
DSC + U-Net	0.979	0.991	0.982	0.993
MCF + atrous + DSC + U-Net (Ours)	**0.981**	**0.992**	**0.983**	**0.994**
Dataset: DRIVE				
U-Net [33]	0.8142	0.9531	0.7537	**0.9820**
MCF + U-Net	0.8168	0.9538	0.7898	0.9790
MCF + atrous + U-Net	0.8230	0.9546	**0.8075**	0.9773
DSC + U-Net	0.8165	0.9527	0.8054	0.9754
MCF + atrous + DSC + U-Net (Ours)	**0.8254**	**0.9561**	0.7934	0.9812

Table 7
Comparison results of model parameters and FLOPs on each component.The FLOPs are calculated based on input image size 448×448.

Methods	Parameters (M)	FLOPs (G)
U-Net [33]	31.04	167.56
MCF + U-Net	37.89	187.50
MCF + atrous + U-Net	23.21	164.48
DSC + U-Net	31.34	169.36
MCF + atrous + DSC + U-Net (Ours)	23.50	166.28

FLOPs are increased from 31.04 M and 167.56G to 37.89 M and 187.50G respectively, as shown in Table 7. After adopting atrous convolution in MCF module (**MCF + atrous + U-Net**), a competitive segmentation performance is still achieved on three datasets, while the model parameters and FLOPs are reduced to 23.21 M and 164.48G respectively.

(2) Ablation Study for Dense Skip Connections

Skip connection is an ingenious idea in original U-Net, which preserves the spatial information that gets lost during the encoding. However, some details like position and boundary information are still lost in the decoding process. Here, we propose dense skip connections, which can fuse low-level details from upper encoder layers and high-level semantics from symmetric encoder feature maps. Table 6 shows the effectiveness of dense skip connections on different segmentation tasks. Compared with U-Net, especially, on the ISIC 2018 and Lung datasets,after adopting dense skip connections (**DSC + U-Net**), all the metrics are improved. However, the improvement of segmentation performance comes at the cost of high model parameters and FLOPs. As shown in Table 7, after introducing dense skip connections (**DSC + U-Net**), the number of model parameters and FLOPs are increased from 31.04 M and 167.56G to 31.34 M and 169.36G respectively.

(3) Ablation Study for CA-Net

Based on the analysis of the efficacy of each component, our CA-Net combines the MCF module and dense skip connections, while introducing the atrous convolution in MCF (**MCF + atrous + DSC + U-Net**), aiming at boosting the performance of different medical segmentation tasks with less parameters and FLOPs. Table 6 shows that, compared to the state-of-the-art U-Net, CA-Net improves all the metrics on the three datasets, except for the *Spec* on DRIVE

dataset. Furthermore, as shown in Table 7, the number of model parameters and FLOPs are reduced from 31.04 M and 167.56G to 23.50 M and 166.28G respectively.

5. Conclusions

In this work, we propose multi-scale CA-Net for multimodal 2D medical image segmentation. First, we employ MCF module after the last convolutional layer of the encoder, which captures multi-scale spatial information through the SCF block and automatically learns channel-wise feature representation through SE block. Then, we adopt dense skip connections to fuse low-level details such as position and boundary of the organs or lesions in medical image and high-level semantics from feature maps. Experimental results demonstrate that, the proposed CA-Net that does not require any pre-processing or post-processing steps outperforms state-of-the-art U-Net and other variants on three public benchmark datasets. Moreover, comparing to U-Net, CA-Net requires less model parameters and FLOPs. We believe that the proposed network is also suitable for other 2D medical image segmentation tasks. In the future work, we would like to add weight information to dense skip connections, making the network automatically learns to focus on different levels of encoder feature maps, so as to improve the feature expression ability of the network.

CRediT authorship contribution statement

Xue Wang: Conceptualization, Methodology, Software, Validation, Writing - original draft, Writing - review & editing. **Zhanshan Li:** Conceptualization, Methodology, Writing - review & editing. **Yongping Huang:** Methodology, Resources, Writing - review & editing, Project administration, Funding acquisition. **Yingying Jiao:** Methodology, Writing - review & editing.

Declaration of Competing Interest

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

Acknowledgements

This research is supported by the National Key Research and Development Program of China (2018YFB0804202, 2018YFB0804203), Regional Joint Fund of NSFC (U19A2057), the National Natural Science Foundation of China (61876070), and the Jilin Province Science and Technology Development Plan Project (20190303134SF).

References

[1] URL:https://www.kaggle.com/kmader/finding-lungs-in-ct-data/data. .
[2] Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. CoRR abs/1802.06955. URL:http://arxiv.org/abs/1802.06955, arXiv:1802.06955. .
[3] Azad, R., Asadi-Aghbolaghi, M., Fathy, M., Escalera, S., 2019. Bi-directional convlstm u-net with densley connected convolutions, in: 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27–28, 2019, IEEE. pp. 406–415. URL: https://doi.org/10.1109/ICCVW.2019.00052, doi: 10.1109/ICCVW.2019.00052. .
[4] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.
[5] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021. Transunet: Transformers make strong encoders for medical image segmentation. CoRR abs/2102.04306. URL:https://arxiv.org/abs/2102.04306, arXiv:2102.04306. .
[6] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (2018) 834–848, https://doi.org/10.1109/TPAMI.2017.2699184.
[7] Chen, L., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. CoRR abs/1706.05587. URL: http://arxiv.org/abs/1706.05587, arXiv:1706.05587. .
[8] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII, Springer. pp. 833–851. URL:https://doi.org/10.1007/978-3-030-01234-2_49, doi: 10.1007/978-3-030-01234-2_49. .
[9] Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N.K., Kittler, H., Halpern, A., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC), in: 15th IEEE International Symposium on Biomedical Imaging, ISBI 2018, Washington, DC, USA, April 4–7, 2018, IEEE. pp. 168–172. URL:https://doi.org/10.1109/ISBI.2018.8363547, doi: 10.1109/ISBI.2018.8363547. .
[10] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H., 2019. Dual attention network for scene segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE. pp. 3146–3154. URL:http://openaccess.thecvf.com/content_CVPR_2019/html/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.html, doi: 10.1109/CVPR.2019.00326. .
[11] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware CNN model, in: 2015 IEEE International Conference on Computer Vision, ICCV 2015, IEEE Computer Society, 2015, pp. 1134–1142, https://doi.org/10.1109/ICCV.2015.135.
[12] K.B. Girum, G. Créhange, A. Lalande, Learning with context feedback loop for robust medical image segmentation, IEEE Trans. Medical Imaging 40 (2021) 1542–1554, https://doi.org/10.1109/TMI.2021.3060497.
[13] M. Gridach, Pydinet: Pyramid dilated network for medical image segmentation, Neural Networks 140 (2021) 274–281, https://doi.org/10.1016/j.neunet.2021.03.023.
[14] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, Ce-net: Context encoder network for 2d medical image segmentation, IEEE Trans. Medical Imaging 38 (2019) 2281–2292, https://doi.org/10.1109/TMI.2019.2903562.
[15] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society. pp. 770–778. URL:https://doi.org/10.1109/CVPR.2016.90, doi: 10.1109/CVPR.2016.90. .
[16] He, X., Yang, S., Li, G., Li, H., Chang, H., Yu, Y., 2019. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press. pp. 8417–8424. URL:https://doi.org/10.1609/aaai.v33i01.33018417, doi: 10.1609/aaai.v33i01.33018417. .
[17] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, A real-time algorithm for signal analysis with the help of the wavelet transform, in: Wavelets, Springer, 1990, pp. 286–297.
[18] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society. pp. 7132–7141. URL: http://openaccess.thecvf.com/content_cvpr_2018/html/Hu_Squeeze-and-Excitation_Networks_CVPR_2018_paper.html, doi: 10.1109/CVPR.2018.00745. .
[19] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, IEEE. pp. 603–612. doi: 10.1109/ICCV.2019.00069. .
[20] S. Hussain, S.M. Anwar, M. Majid, Segmentation of glioma tumors in brain using deep convolutional neural network, Neurocomputing 282 (2018) 248–261, https://doi.org/10.1016/j.neucom.2017.12.032.
[21] N. Ibtehaz, M.S. Rahman, Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation, Neural Networks 121 (2020) 74–87, https://doi.org/10.1016/j.neunet.2019.08.025.
[22] A. Javed, Y. Kim, M.C.K. Khoo, S.L.D. Ward, K.S. Nayak, Dynamic 3-d MR visualization and detection of upper airway obstruction during sleep using region-growing segmentation, IEEE Trans. Biomed. Eng. 63 (2016) 431–437, https://doi.org/10.1109/TBME.2015.2462750.
[23] Q. Li, B. Feng, L. Xie, P. Liang, H. Zhang, T. Wang, A cross-modality learning approach for vessel segmentation in retinal images, IEEE Trans. Medical Imaging 35 (2016) 109–118, https://doi.org/10.1109/TMI.2015.2457891.
[24] L. Liu, J. Cheng, Q. Quan, F. Wu, Y. Wang, J. Wang, A survey on u-shaped networks in medical image segmentations, Neurocomputing 409 (2020) 244–258, https://doi.org/10.1016/j.neucom.2020.05.070.

[25] Y. Liu, H. Chen, X. Shen, Y. Huang, Gamma correction FCM algorithm with conditional spatial information for image segmentation. KSII Trans. Internet, Inf. Syst. 12 (2018) 4336–4354, https://doi.org/10.3837/tiis.2018.09.012.
[26] Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., Wang, X., 2021a. Deep dual consecutive network for human pose estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021, Computer Vision Foundation/ IEEE. pp. 525–534. URL:https://openaccess.thecvf.com/content/CVPR2021/html/Liu_Deep_Dual_Consecutive_Network_for_Human_Pose_Estimation_CVPR_2021_paper.html. .
[27] Liu, Z., Qian, P., Wang, X., Zhuang, Y., Qiu, L., Wang, X., 2021b. Combining graph neural networks with expert knowledge for smart contract vulnerability detection. CoRR abs/2107.11598. URL:https://arxiv.org/abs/2107.11598, arXiv:2107.11598. .
[28] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society. pp. 3431–3440. doi: 10.1109/CVPR.2015.7298965. .
[29] F. Milletari, N. Navab, S. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: Fourth International Conference on 3D Vision, 3DV 2016,,, IEEE Computer Society, Stanford, CA, USA, 2016, pp. 565–571, https://doi.org/10.1109/3DV.2016.79.
[30] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention u-net: Learning where to look for the pancreas. CoRR abs/1804.03999. URL:http://arxiv.org/abs/1804.03999, arXiv:1804.03999. .
[31] T.X. Pham, P. Siarry, H. Oulhadj, Integrating fuzzy entropy clustering with an improved PSO for MRI brain image segmentation, Appl. Soft Comput. 65 (2018) 230–242, https://doi.org/10.1016/j.asoc.2018.01.003.
[32] Ren, S., He, K., Girshick, R.B., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks, in: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal, Quebec, Canada, pp. 91–99. URL:https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html. .
[33] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015–18th International Conference Munich, Germany, October 5–9, 2015, Proceedings, Part III, Springer. pp. 234–241. URL:https://doi.org/10.1007/978-3-319-24574-4_28, doi: 10.1007/978-3-319-24574-4_28. .
[34] R. Rouhi, M. Jafari, S. Kasaei, P. Keshavarzian, Benign and malignant breast tumors classification based on region growing and CNN segmentation, Expert Syst. Appl. 42 (2015) 990–1002, https://doi.org/10.1016/j.eswa.2014.09.020.
[35] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi, M.C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama, P. Cazzaniga, Use-net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional MRI datasets, Neurocomputing 365 (2019) 31–43, https://doi.org/10.1016/j.neucom.2019.07.006.
[36] Y. Shu, J. Zhang, B. Xiao, W. Li, Medical image segmentation based on active fusion-transduction of multi-stream features, Knowl. Based Syst. 220 (2021) , https://doi.org/10.1016/j.knosys.2021.106950 106950.
[37] G. Song, K.M. Lee, Bi-directional seed attention network for interactive image segmentation, IEEE Signal Process. Lett. 27 (2020) 1540–1544, https://doi.org/10.1109/LSP.2020.3019970.
[38] Song, H., Wang, W., Zhao, S., Shen, J., Lam, K., 2018. Pyramid dilated deeper convlstm for video salient object detection, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision - ECCV 2018–15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XI, Springer. pp. 744–760. URL:https://doi.org/10.1007/978-3-030-01252-6_44, doi: 10.1007/978-3-030-01252-6_44. .
[39] J. Staal, M.D. Abràmoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, Ridge-based vessel segmentation in color images of the retina, IEEE Trans. Medical Imaging 23 (2004) 501–509, https://doi.org/10.1109/TMI.2004.825627.
[40] J. Sun, Y. Peng, Y. Guo, D. Li, Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3d FCN, Neurocomputing 423 (2021) 34–45, https://doi.org/10.1016/j.neucom.2020.10.031.
[41] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Singh, S.P., Markovitch, S. (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4–9, 2017, San Francisco, California, USA, AAAI Press. pp. 4278–4284. URL:http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806. .
[42] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, IEEE Computer Society. pp. 1–9. doi: 10.1109/CVPR.2015.7298594. .
[43] Z. Tian, L. Liu, Z. Zhang, B. Fei, Superpixel-based segmentation for 3d prostate MR images, IEEE Trans. Medical Imaging 35 (2016) 791–801, https://doi.org/10.1109/TMI.2015.2496296.
[44] Tseng, K., Lin, Y., Hsu, W.H., Huang, C., 2017. Joint sequence learning and cross-modality convolution for 3d biomedical segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society. pp. 3739–3746. URL:https://doi.org/10.1109/CVPR.2017.398, doi: 10.1109/CVPR.2017.398. .
[45] G. Wu, J. Han, Z. Lin, G. Ding, B. Zhang, Q. Ni, Joint image-text hashing for fast large-scale cross-media retrieval using self-supervised deep learning, IEEE Trans. Ind. Electron. 66 (2019) 9868–9877, https://doi.org/10.1109/TIE.2018.2873547.
[46] L. Xie, Q. Yu, Y. Zhou, Y. Wang, E.K. Fishman, A.L. Yuille, Recurrent saliency transformation network for tiny target segmentation in abdominal CT scans, IEEE Trans. Medical Imaging 39 (2020) 514–525, https://doi.org/10.1109/TMI.2019.2930679.
[47] Z. Yan, X. Yang, K. Cheng, A three-stage deep learning model for accurate retinal vessel segmentation, IEEE J. Biomed. Health Informatics 23 (2019) 1427–1436, https://doi.org/10.1109/JBHI.2018.2872813.
[48] Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., Chua, T., 2020a. Tree-augmented cross-modal encoding for complex-query video retrieval, in: Huang, J., Chang, Y., Cheng, X., Kamps, J., Murdock, V., Wen, J., Liu, Y. (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020, ACM. pp. 1339–1348. doi: 10.1145/3397271.3401151. .
[49] Yang, X., Liu, X., Jian, M., Gao, X., Wang, M., 2020b. Weakly-supervised video object grounding by exploring spatio-temporal contexts, in: Chen, C.W., Cucchiara, R., Hua, X., Qi, G., Ricci, E., Zhang, Z., Zimmermann, R. (Eds.), MM '20: The 28th ACM International Conference on Multimedia, Virtual Event/ Seattle, WA, USA, October 12–16, 2020, ACM. pp. 1939–1947. doi: 10.1145/3394171.3413610. .
[50] Q. Yu, Y. Gao, Y. Zheng, J. Zhu, Y. Dai, Y. Shi, Crossover-net: Leveraging vertical-horizontal crossover relation for robust medical image segmentation, Pattern Recognit. 113 (2021) , https://doi.org/10.1016/j.patcog.2020.107756 107756.
[51] Yu, Q., Xie, L., Wang, Y., Zhou, Y., Fishman, E.K., Yuille, A.L., 2018. Recurrent saliency transformation network: Incorporating multi-stage visual cues for small organ segmentation, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018, IEEE Computer Society. pp. 8280–8289. URL:http://openaccess.thecvf.com/content_cvpr_2018/html/Yu_Recurrent_Saliency_Transformation_CVPR_2018_paper.html, doi: 10.1109/CVPR.2018.00864. .
[52] J. Zhang, Y. Xie, Y. Wang, Y. Xia, Inter-slice context residual learning for 3d medical image segmentation, IEEE Trans. Medical Imaging 40 (2021) 661–672, https://doi.org/10.1109/TMI.2020.3034995.
[53] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017, IEEE Computer Society. pp. 6230–6239. URL:https://doi.org/10.1109/CVPR.2017.660, doi: 10.1109/CVPR.2017.660. .
[54] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, https://doi.org/10.1016/j.aiopen.2021.01.001.
[55] S. Zhou, J. Wang, M. Zhang, Q. Cai, Y. Gong, Correntropy-based level set method for medical image segmentation and bias correction, Neurocomputing 234 (2017) 216–229, https://doi.org/10.1016/j.neucom.2017.01.013.
[56] S. Zhou, J. Wang, S. Zhang, Y. Liang, Y. Gong, Active contour model based on local and global intensity information for medical image segmentation, Neurocomputing 186 (2016) 107–118, https://doi.org/10.1016/j.neucom.2015.12.073.
[57] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: Redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Medical Imaging 39 (2020) 1856–1867, https://doi.org/10.1109/TMI.2019.2959609.

Xue Wang received the M.S. degree from Jilin University of China in 2010. She is currently a Ph.D. student at the College of Computer Science and Technology, Jilin University. Her research interests include machine learning and visual reasoning, and image processing.

Zhanshan Li is a professor at Jilin University. His interests include constraint optimization and constraintsolving, machine learning.

Yongping Huang is a professor at Jilin University. His interests include embedded software architecture, machine learning, edge computing, and complex adaptive control system.

Yingying Jiao was a research assistant with the School of Computing, National University of Singapore, Singapore. She is currently pursuing the Ph.D. degree in the Department of Computer Science and Technology, Jilin University, China. Her research interests include deep learning, image processing, and blockchain technology.