Computers in Biology and Medicine 154 (2023) 106626

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

DCSAU-Net: A deeper and more compact split-attention U-Net for medical image segmentation

Qing Xu [a,*], Zhicheng Ma [b], Na HE [c], Wenting Duan [a]

[a] The School of Computer Science, University of Lincoln, Lincolnshire, LN6 7TS, United Kingdom
[b] The College of Computer Science and Technology, Zhejiang Gongshang University, Zhejiang, 310018, China
[c] The Sino-German Institute of Design and Communication, Zhejiang Wanli University, Zhejiang, 315100, China

ARTICLE INFO

Keywords:
Medical image segmentation
Multi-scale fusion attention
Depthwise separable convolution
Computer-aided diagnosis

ABSTRACT

Deep learning architecture with convolutional neural network achieves outstanding success in the field of computer vision. Where U-Net has made a great breakthrough in biomedical image segmentation and has been widely applied in a wide range of practical scenarios. However, the equal design of every downsampling layer in the encoder part and simply stacked convolutions do not allow U-Net to extract sufficient information of features from different depths. The increasing complexity of medical images brings new challenges to the existing methods. In this paper, we propose a deeper and more compact split-attention u-shape network, which efficiently utilises low-level and high-level semantic information based on two frameworks: primary feature conservation and compact split-attention block. We evaluate the proposed model on CVC-ClinicDB, 2018 Data Science Bowl, ISIC-2018, SegPC-2021 and BraTS-2021 datasets. As a result, our proposed model displays better performance than other state-of-the-art methods in terms of the mean intersection over union and dice coefficient. More significantly, the proposed model demonstrates excellent segmentation performance on challenging images. The code for our work and more technical details can be found at https://github.com/xq141839/DCSAU-Net.

1. Introduction

Common types of cancer such as colon cancer, multiple myeloma and melanoma, are still one of the major causes of human suffering and death globally. Medical image analysis plays an essential role in terms of diagnosing and treating these diseases [1]. For example, numerous cells in a microscopy image are able to illustrate the stage of diseases, assist in discriminating tumour types, support in insight of cellular and molecular genetic mechanisms, and present valuable information to many other applications, such as cancer and chronic obstructive pulmonary disease [2]. Traditionally, medical images are analysed by pathologists manually. In other words, the result of diagnosis is usually dominated by the experience of medical experts, which can be time-consuming, subjective, and error-prone [3]. Computer-aided diagnosis (CAD) has received significant attention from both pathological researchers and clinical practice, which is mainly depend on the result of medical image segmentation [4]. Different from classification and detection tasks, the target of biomedical image segmentation is to separate the specified object from the background in an image, which is able to provide patients with more detailed disease analysis [5]. Existing classic segmentation algorithms are based on edge detection, thresholding, morphology, distances between two objects and pixel energy, such as Otsu thresholding [6], Snake [7] and Fuzzy algorithms [8]. Each algorithm has its own parameters to accommodate different requirements. However, these algorithms often show limited performance on the generalisation of complex datasets [9]. The segmentation performance of these methods is also affected by image acquisition quality. For example, some pathological images may be blurred or contain noises. Other situations could have negative influences too, including uneven illumination, low image contrast between foreground and background, and complex tissue background [10]. Therefore, it is essential to construct a powerful and generic model which can achieve adequate robustness on challenging images and work for different biomedical applications.

Convolutional neural network (CNN) based encoder–decoder architectures have outperformed traditional image processing methods in various medical image segmentation tasks [11]. The success of these models is largely due to the skip connection strategy that incorporates the low-level semantic information with high-level semantic information to generate the final mask [12]. However, many improved architectures only focus on optimising algorithms in terms of in-depth

* Corresponding author.
E-mail address: xq14183925@gmail.com (Q. Xu).

https://doi.org/10.1016/j.compbiomed.2023.106626
Received 23 September 2022; Received in revised form 13 January 2023; Accepted 28 January 2023
Available online 1 February 2023

feature extraction, which ignores the loss of high-resolution information in the header of the encoder. The sufficient feature maps extracted from this layer is able to help to compensate for the spatial information lost during the pooling operations [13].

In this paper, we propose an encoder–decoder architecture for medical image segmentation, called DCSAU-Net. In the encoder part, our model first adopts a primary feature conservation (PFC) strategy that reduces the number of parameters, amount of computation and integrates the long-range spatial information of the network in the low-level semantic layer. The rich primary feature obtained from this layer will be delivered to our constructed module: compact split-attention (CSA) block. The CSA module strengthens the feature representation of different channels using a multi-path attention structure. Each path contains a different number of convolutions so that the CSA module can output mixed feature maps with different receptive field scales. Both new frameworks are designed with residual style in order to alleviate gradient vanishing problem with increasing layers. For the decoder, encoded features in every downsampling layer are concatenated with corresponding upsampled features by skip connection. we apply the same CSA block to complete efficient feature extraction from the combined features. The proposed DCSAU-Net is easy to train without any extra support samples (e.g. initialised mask or edge). The main contributions of this work can be summarised as follows:

(1) A mechanism, PFC, is embedded in our DCSAU-Net to capture sufficient primary features from the input images. Compared with other common designs, PFC not only improves computational efficiency but also extends the receptive field of the network.
(2) To enhance the multi-scale representation of DCSAU-Net, we build a CSA block that adopts multi-branch feature groups with attention mechanism. Each group is comprised of a different number of convolutions in order to output feature maps with the combination of different receptive field sizes.
(3) Experimental analysis is conducted with five different medical image segmentation datasets, including 2018 Data Science Bowl [14], ISIC-2018 Lesion Boundary Segmentation [15,16], CVC-ClinicDB [17], and two multi-class segmentation datasets: SegPC-2021 [18] and BraTS-2021 [19–21]. Evaluation results demonstrate that our proposed DCSAU-Net shows better performance than other state-of-the-art (SOTA) segmentation methods in terms of standard computer vision metrics — intersection over union and dice coefficient, which can be a new SOTA method for medical image segmentation.

2. Related work

2.1. Medical image segmentation

Deep learning methods based on convolutional neural network (CNN) have indicated outstanding performance in medical image segmentation. U-Net, proposed by Ronneberger et al. [22], is comprised of two components: encoder and decoder. Upsampling operators are added in the decoder, which is used to recover the resolution of input images. Also, features extracted from the encoder are combined with upsampled results to achieve precise localisation. U-Net shows a favourable segmentation performance in different kinds of medical images. Inspired by this architecture, Zhou et al. [23] presented a nested U-Net (Unet++) for medical image segmentation. To reduce the semantic information loss of feature fusion between encoder and decoder, a series of nested and skip pathways are added to the model. Huang et al. [24] designed another full-scale skip connection method that combines low-resolution information and high-resolution information in different scales. Jha et al. [25] constructed a DoubleU-Net network that organises two U-Net architectures sequentially. In the encoder part, Atrous Spatial Pyramid Pooling (ASPP) is constructed at the end of each downsample layer to obtain contextual information. The evaluation result demonstrates that DoubleU-Net performs well in polyp, lesion boundary and nuclei segmentation. The gradient vanishing issue has been discovered when trying to converge deeper networks. To address this problem, He et al. [26] introduced a deep residual architecture (ResNet) that had been widely applied in different segmentation networks. For medical image segmentation, Jha et al. [27] constructed an advanced u-shape architecture for polyp segmentation, called ResUNet++. This model involves residual style, squeeze and excitation module, ASPP, and attention mechanism. Tarasiewicz et al. [28] developed lightweight U-Nets for brain tumour segmentation. The inception modules and dense blocks are embedded in the encoder and decoder for wide spatial information collection. However, most of the models are only evaluated on a single dataset or binary segmentation tasks. In contrast to modifying the architecture of the network, Isensee et al. [29] proposed nnU-Net that emphasises the importance of data pre-processing, model training strategy and result inference. For any task, nn-Unet is able to automatically search the most optimal parameters of the model. It has been evaluated on 23 public datasets with minor-change U-Nets and has shown outstanding performance compared to other algorithms. On the contrary, our approach focuses on the enhancement of the network itself as these techniques can be also applied to any architecture and improve its performance [30,31].

2.2. Attention mechanisms

In previous years, the attention mechanism has rapidly appeared in computer vision. SENet [32], one of channel attention, has been widely applied in medical image segmentation [33,34]. It uses a squeeze module, with global average pooling, to collect global spatial information, and an excitation module to obtain the relationship between each channel in feature maps. Spatial attention can be referred to as an adaptive spatial location selection mechanism. For instance, Oktay et al. [35] introduced an attention U-Net using a bottom-up attention gate, which can precisely focus on a specific region that highlights useful features without extra computational costs and model parameters. Furthermore, transformer [36] has received a lot of attention recently because its success in natural language processing (NLP). Dosovitskiy et al. [37] developed a vision transformer (ViT) architecture for computer vision tasks and indicated comparable performance to CNN. Also, a series of ungraded ViT has been in a wider range of fields. Xu et al. [38] proposed LeViT-UNet to collect distant spatial information from features. In addition, transformer has demonstrated strong performance when incorporated with CNN. Chen et al. [39], provided TransUNet that selects CNN as the first half of the encoder to obtain image patches and uses the transformer model to extract the global contexts. The final mixed feature in the decoder can achieve more accurate localisation. However, transformer-based networks usually contain a large number of parameters and consume more computing sources. We optimise existing attention architectures and propose a lightweight attention module.

2.3. Depthwise separable convolution

Depthwise separable convolution is an efficient neural network architecture proposed by Howard et al. [40]. Each convolution filter in this architecture is responsible for one input channel. Compared with a standard convolution, depthwise convolution not only can achieve the same effects but also costs fewer number of parameters and computations. However, it only extracts features of every input channel. To combine the information between the channels and create new feature maps, a 1×1 convolution, called pointwise convolution, follows a depthwise convolution. The final MobileNets model was established and considered as a new backbone in deep learning. In the image classification task, Chollet [41] used depthwise separable convolution

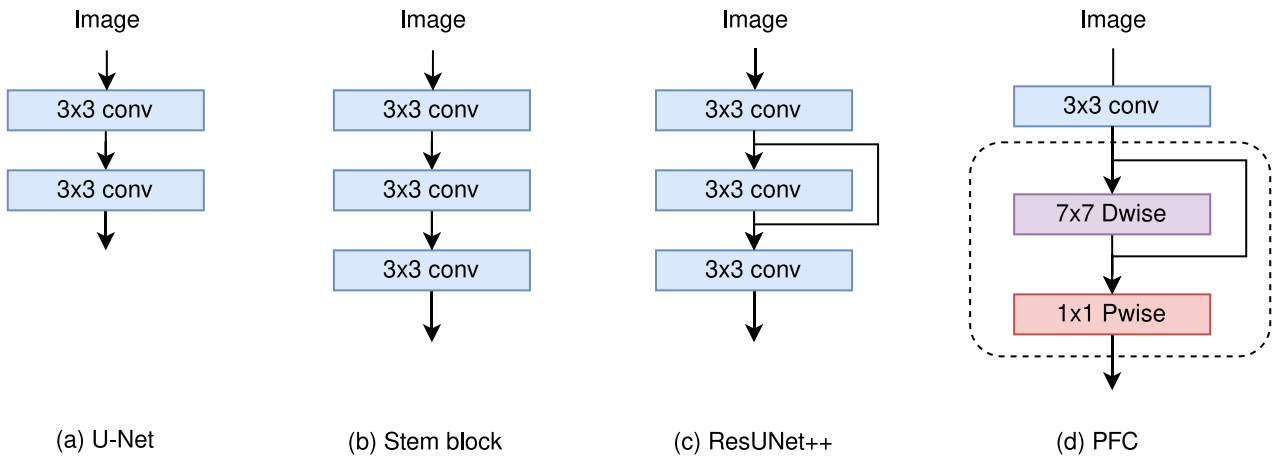

Fig. 1. Comparing our PFC strategy with U-Net [22], Stem block [44] and ResUNet++ [27] designs used to extract the low-level semantic information from the input images.

to construct an Xception model that outperformed previous SOTA methods and showed lower complexity. However, Sandler et al. [42] observed that depthwise convolution performs poorly in the low-channel feature maps. To tackle aforementioned issues, they proposed a new MobileNetV2 model that adds a 1×1 convolution in front of the depthwise convolution in order to increase the dimension of features in advance. Compared with MobileNets, MobileNetV2 does not raise the number of parameters but decreases the degradation of performance. In medical image segmentation, Qi et al. [43] introduced an X-net model for 3D brain stroke lesion segmentation. A feature similarity module (FSM) was created to capture distance spatial information in feature maps using depthwise separable convolution. The experiment results demonstrate the X-net model costs only half the number of parameters of other SOTA models to achieve higher performance.

3. Method

3.1. Primary feature conservation

For most of medical image segmentation networks, the convolutions used in the first downsampling block operation is to extract low-level semantic information from images. The U-Net architecture [22] in Fig. 1(a) has been widely used in different models [25,35]. The stem block [44] in Fig. 1(b) is usually designed to obtain the same receptive field as 7×7 convolution and reduce the number of parameters. The first feature scale downsampling layer in ResUNet++ [27] adds skip connection strategy to mitigate the potential impact of the gradient vanish, which is shown in Fig. 1(c). Although stacking more convolutional blocks can extend the receptive field of neural network, the number of parameters and the amount of computation will increase rapidly. The stability of the model may be destroyed. Also, recent research suggests that the valid receptive field will decrease to some extent when the number of stacked 3×3 convolutions keep increasing [45]. To address this issue, we introduce a new primary feature conservation (PFC) strategy in the first downsampling block, which is provided in Fig. 1(d). The main refinement of our module adopts depthwise separable convolution, consisting of 7×7 depthwise convolution followed by 1×1 pointwise convolution. As depthwise separable convolution decreases the costs of computation and the number of parameters compared to the standard convolution [40], we have an opportunity to apply large kernel sizes on the depthwise convolution in order to merge distant spatial information and preserve primary features as much as possible in the low-level semantic layer. The 1×1 pointwise convolution is used to combine channel information. Also, 3×3 convolution is added to the head of this module for raising the feature channel because depthwise separable convolution shows the degradation of performance on low-dimensional features [42]. Every convolution is followed by a ReLU activation and BatchNorm. To avoid gradient vanish, the PFC block is constructed with residual style. To this end, our proposed PFC module can improve performance without increasing the number of parameters and computational costs. In addition, the reason for using depthwise convolution with 7×7 kernel size will be explained in Section 5.

3.2. Compact split-attention block

The VGGNet [46] and typical residual structures [26] have been applied in many previous semantic segmentation networks, such as DoubleUnet [25] and ResUnet [47]. However, convolutional layers in VGGNet are stacked directly, which means every feature layer has a comparatively constant receptive field [48]. In medical image segmentation, different lesions may have different sizes. Sufficient representation of multi-scale features is beneficial for the model to perceive data features. Recently various models learning the representation via cross-channel features have been proposed, such as ResNeSt [49]. Inspired by these methods, we develop a new compact split-attention (CSA) architecture.

An overview of CSA block is illustrated in Fig. 2. The ResNeSt utilises large channel-split groups for feature extraction, which is more efficient for general computer vision tasks with adequate data and costs massive parameters. Furthermore, each group of this model adopts the same convolutional operations that receive an equal receptive field size. To optimise the structure and make it more suitable for medical image segmentation, our proposed block maintains two feature groups ($N = 2$) to reduce the number of parameters in the entire network. These two groups split from the input features will be fed into different transformations F_i. Both two groups involve one 1×1 convolution followed by one 3×3 convolution. To improve the representation across channels, the output feature maps of the other group (F_2) will sum the result of the first group (F_1) and go through another 3×3 convolution, which can receive semantic information from both split groups and expand the receptive field of the network. Therefore, the CSA block presents a stronger ability to extract both global and local information from feature maps. Mathematically, the fusion feature maps can be defined as:

$$\hat{U} = \sum_{i=1}^{N} F_i(X_i), \quad \hat{U} \in R^{H \times W \times C} \tag{1}$$

Where H, W and C are the scales of output feature maps. The channel-wise statistics generated by global average pooling collect global spatial information, which is produced by compressing transformation output through spatial dimensions and the cth component calculated by:

$$S_c = \frac{1}{H \times W} \sum_{\alpha=1}^{H} \sum_{\beta=1}^{W} \hat{U}_c(\alpha, \beta), \quad S \in R^C \tag{2}$$

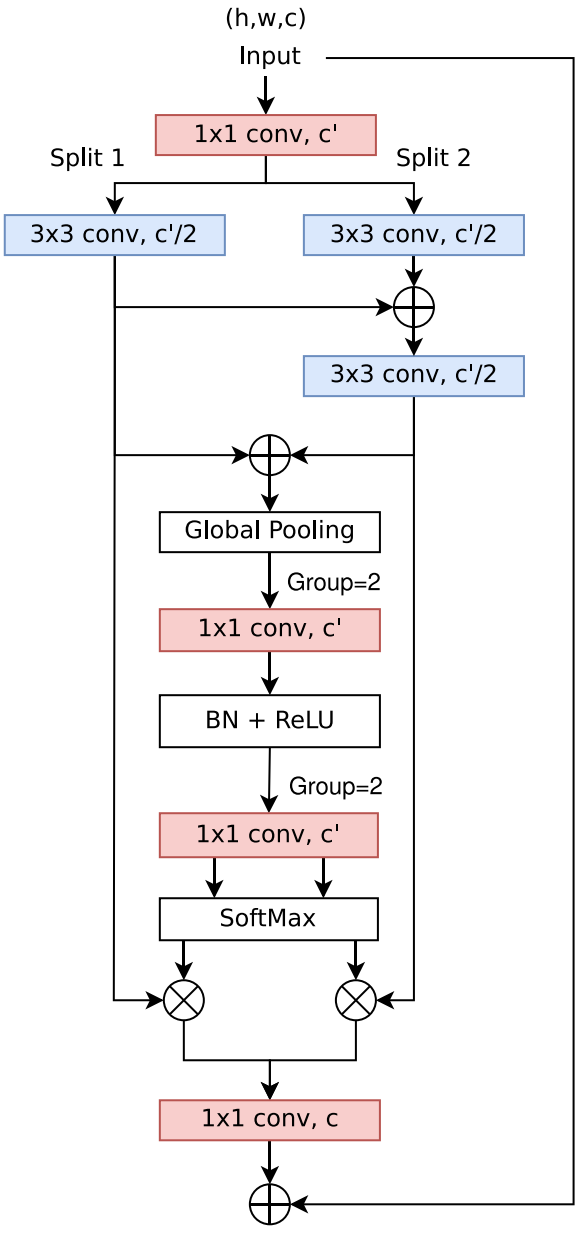

Fig. 2. The framework of CSA block.

The channel-wise soft attention is used for aggregating a weighted fusion represented by cardinal group representation, where a split weighted combination can catch crucial information in feature maps. Then the cth channel of feature maps is calculated as:

$$V_c = \sum_{i=1}^{N} a_i(c)\, F_i(X_i) \tag{3}$$

Where a_i is a (soft) assignment weight designed by:

$$a_i(c) = \frac{exp(\mathcal{G}_i^c(S))}{\sum_{j=1}^{N} exp(\mathcal{G}_j^c(S))} \tag{4}$$

Here $\mathcal{G}_i^c$ indicates the weight of global spatial information S to the cth channel and is quantified using two 1×1 convolutions with BatchNorm and ReLU activation. As a result, the full CSA block is designed with a standard residual architecture that the output Y is calculated using a skip connection: $Y = V + X$, when the shape of output feature maps is the same as the input feature maps. Otherwise, an extra transformation T will be applied on the skip connection to obtain the same shape. For instance, T can be convolution with a stride or mix of convolution and pooling.

3.3. DCSAU-Net architecture

For medical image segmentation, we establish a model using the proposed PFC strategy and CSA block following the encoder–decoder architecture, which is referred to as DCSAU-Net and shown in Fig. 3. The encoder of DCSAU-Net first uses PFC strategy to extract low-level semantic information from the input images. The depthwise separable convolution with a large 7×7 kernel size is able to broaden the receptive field of the network and preserve primary features without increasing the number of parameters. The CSA block applies multi-path feature groups with a different number of convolutions and the attention mechanism, which incorporates channel information across different receptive field scales and highlights meaningful semantic features. Each of block is followed by a 2×2 max pooling with stride 2 for performing a downsampling operation. Every decoder sub-network starts with an upsampling operator to recover the original size of the input image step by step. The skip connections are used to concatenate these feature maps with the feature maps from the corresponding encoder layer, which mixes low-level and high-level semantic information to generate a precise mask. The skip connections are followed by CSA blocks to alleviate the gradient vanishing problem and capture efficient features. Finally, a 1×1 convolution succeeded by a sigmoid or softmax layer is used to output the binary or multi-class segmentation mask.

4. Experiments and results

4.1. Datasets

To evaluate the effectiveness of DCSAU-Net, we test it on five publicly available medical image datasets.

- CVC-ClinicDB [17] is a frequently-used dataset for the polyp segmentation task. It is also the training database for the MICCAI 2015 Sub-Challenge on Automatic Polyp Detection Challenge.
- The second dataset used in this study is from the 2018 Data Science Bowl challenge [14], which is used for the nuclei segmentation task. The dataset labels every cell in microscopic images.
- Another dataset used in our experiment is from a sub-task in the ISIC-2018 challenge [15,16]. The target of training the dataset is to develop a model for lesion boundary segmentation.
- In order to assess the performance of the proposed architecture on the multi-class segmentation task, we add the SegPC-2021 dataset [18] in our experiment. Each of image in the dataset includes two different Myeloma Plasma cells.
- BraTS-2021 dataset [19–21] is used to evaluate the performance of models for brain tumour segmentation. It involves three different labels: GD-enhancing tumour, invaded tissue and necrotic tumour core.

More details about dataset organisation are presented in Table 1. We use a fixed random seed for all data split, which has been included in our code. All of these datasets are related to clinic diagnosis. Therefore, their segmentation result can be significant for patients.

4.2. Evaluation metrics

Mean intersection over union (mIoU), accuracy, recall, precision and dice coefficient (DSC) are standard metrics for medical image segmentation, where mIoU is a common metric used in competitions to compare the performance between each of models. For a more exhaustive comparison between the performance of DCSAU-Net and other popular models, we calculate each of these metrics in our experiment. The p-values of the paired t-tests are also reported.

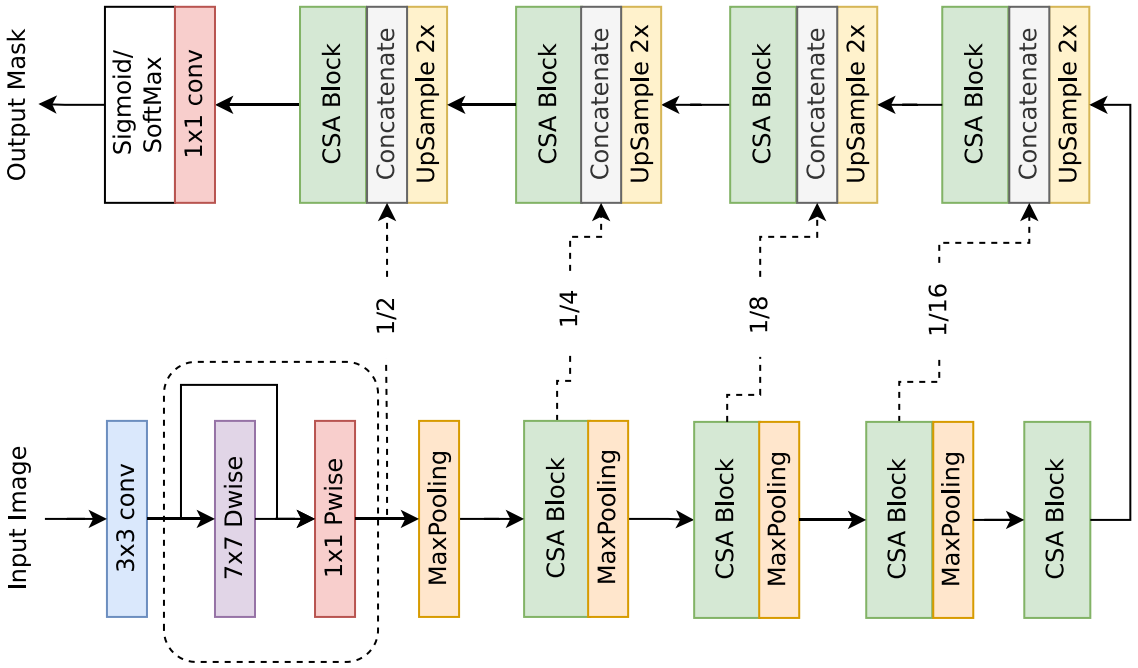

Fig. 3. The presentation of DCSAU-Net with PFC strategy and CSA block.

Table 1
Details of the medical segmentation datasets used in our experiments.

Dataset	Images	Input size	Train	Valid	Test
CVC-ClinicDB	612	384 × 288	441	110	61
2018 Data Science Bowl	670	Variable	483	120	67
ISIC 2018	2594	Variable	1868	467	259
SegPC 2021	498	Variable	360	89	49
BraTS 2021	1251	240 × 240	900	226	125

4.3. Data augmentation

Medical image datasets usually have a limited number of samples to be available in the training phase due to obtaining and annotating images are expensive and time-consuming [50]. Therefore, the model is prone to overfitting. To mitigate this issue, data augmentation methods are generally used in the training stage to extend the diversity of samples and enhance the model generalisation. In our experiment, we randomly apply horizontal flip, rotation and cutout with a probability of 0.25 to the training set of each dataset.

4.4. Implementation details

All experiments are implemented using PyTorch 1.10.0 framework on a single NVIDIA V100 Tensor Core GPU, 8-core CPU and 16 GB RAM. We use a common segmentation loss function, Dice loss, and an Adam optimiser with a learning rate of 1e−4 to train all models. The number of batch sizes and epochs are set to 16 and 200 respectively. During training, we resize the images to 256 × 256 for CVC-ClinicDB and 2018 Data Science Bowl datasets. For ISIC-2018 and SegPC-2021 datasets, the input images are resized to 512 × 512. To train models on BraTS dataset, we extract 2D slides from all multi-parametric magnetic resonance imaging (mpMRI) scans. Also, we apply ReduceLROnPlateau to optimise the learning rate. All experiments on five datasets are conducted on the same train, validation, and test datasets. In addition, we train other SOTA models with default parameters, meanwhile, a pretrained ViT model is loaded when training the TransUNet and LeViT-UNet. The rest of models are trained from scratch.

4.5. Results

In this section, we present quantitative results on five different biomedical image datasets and compare our proposed architecture with other SOTA methods.

4.5.1. Comparison on CVC-ClinicDB dataset

The quantitative results on CVC-ClinicDB dataset are shown in Table 2. For medical image segmentation, the performance of the network on mIoU and DSC metrics usually receives more attention. From Table 2, DCSAU-Net achieves a DSC of 0.916 and a mIoU of 0.861, which outperforms DoubleU-Net by 2.0% in terms of DSC and 2.5% in mIoU. Particularly, our proposed model provides a significant improvement over the two recent transformer-based architectures, where the mIoU of DCSAU-Net is 6.2% and 10.7% higher than TransUNet and LeViT-UNet, and the DSC of DCSAU-Net is 4.9% and 8.8% higher than these two models respectively.

4.5.2. Comparison on SegPC-2021 dataset

For medical image analysis, some of medical images may have multi-class objects that need to be segmented out. To satisfy this demand, we evaluate all models on SegPC-2021 dataset with two different kinds of cells. The quantitative results are provided in Table 3. Compared with other SOTA models, DCSAU-Net displays the best performance in all defined metrics. Specifically, our proposed method produces a mIoU score of 0.8048 with a more significant rise of 3.6% over Unet++ and 2.8% in DSC compared to the DoubleU-Net architecture.

4.5.3. Comparison on 2018 data science bowl dataset

Nuclei segmentation plays an important role in the biomedical image analysis. We use an open-access dataset from 2018 Data Science Bowl challenge to evaluate the performance of DCSAU-Net and other SOTA networks. A comparison between each model is presented in Table 4. The results demonstrate that DCSAU-Net achieves a DSC of 0.914 which is 1.9% higher than TransUNet and mIoU of 0.850, which is 2.5% higher than UNet3+.

Table 2
Results on the CVC-ClinicDB.

Method	Accuracy	Precision	Recall	DSC	mIoU	P-values
U-Net [22]	0.984±0.019	0.882±0.195	0.893±0.176	0.872±0.189	0.809±0.213	3.182e−02
Unet++ [23]	0.984±0.022	**0.919±0.139**	0.859±0.197	0.876±0.184	0.811±0.196	6.961e−03
Attention-UNet [35]	0.986±0.016	0.904±0.170	0.901±0.185	0.895±0.168	0.835±0.179	8.776e−02
ResUNet++ [27]	0.982±0.021	0.870±0.191	0.853±0.213	0.854±0.196	0.781±0.213	5.356e−04
R2U-Net [51]	0.978±0.028	0.880±0.185	0.847±0.223	0.841±0.205	0.765±0.224	2.108e−04
DoubleU-Net [25]	0.986±0.017	0.892±0.179	0.912±0.197	0.896±0.173	0.836±0.196	4.903e−02
UNet3+ [24]	0.984±0.022	0.907±0.152	0.885±0.155	0.892±0.171	0.827±0.191	7.089e−02
TransUNet [39]	0.982±0.209	0.876±0.199	0.873±0.191	0.867±0.188	0.799±0.201	1.571e−03
LeViT-UNet [38]	0.980±0.023	0.849±0.241	0.826±0.232	0.828±0.233	0.754±0.244	1.118e−03
DCSAU-Net	**0.990±0.015**	0.917±0.148	**0.920±0.143**	**0.916±0.141**	**0.861±0.156**	1.000e+00

Table 3
Results on the SegPC 2021 (Multiple myeloma plasma cells segmentation challenge).

Method	Accuracy	Precision	Recall	DSC	mIoU	P-values
U-Net [22]	0.939±0.053	0.842±0.142	0.879±0.118	0.855±0.119	0.766±0.148	2.299e−07
Unet++ [23]	0.942±0.058	0.855±0.142	0.876±0.141	0.857±0.127	0.770±0.163	4.483e−09
Attention-UNet [35]	0.940±0.048	0.845±0.143	0.866±0.125	0.849±0.117	0.757±0.147	7.885e−08
ResUNet++ [27]	0.934±0.051	0.838±0.118	0.858±0.101	0.840±0.086	0.736±0.121	2.248e−08
R2U-Net [51]	0.933±0.056	0.852±0.122	0.831±0.136	0.834±0.112	0.744±0.128	4.955e−09
DoubleU-Net [25]	0.937±0.052	0.833±0.120	0.896±0.084	0.858±0.089	0.763±0.130	7.650e−08
UNet3+ [24]	0.939±0.051	0.848±0.119	0.866±0.078	0.852±0.083	0.766±0.131	4.622e−06
TransUNet [39]	0.939±0.047	0.822±0.130	0.869±0.121	0.838±0.113	0.741±0.146	2.922e−08
LeViT-UNet [38]	0.939±0.049	0.850±0.120	0.837±0.115	0.837±0.101	0.738±0.137	3.540e−07
DCSAU-Net	**0.950±0.045**	**0.871±0.113**	**0.910±0.067**	**0.886±0.078**	**0.806±0.121**	1.000e+00

Table 4
Results on the 2018 Data Science Bowl.

Method	Accuracy	Precision	Recall	DSC	mIoU	P-values
U-Net [22]	0.955±0.047	0.872±0.105	0.920±0.111	0.887±0.090	0.808±0.126	5.039e−05
Unet++ [23]	0.955±0.047	0.874±0.122	0.918±0.141	0.886±0.132	0.814±0.150	5.534e−04
Attention-UNet [35]	0.953±0.046	0.870±0.151	0.918±0.136	0.887±0.134	0.816±0.152	9.075e−03
ResUNet++ [27]	0.954±0.048	0.900±0.120	0.903±0.104	0.894±0.104	0.822±0.138	2.973e−02
R2U-Net [51]	0.956±0.047	0.884±0.135	0.911±0.140	0.891±0.135	0.822±0.156	1.082e−01
DoubleU-Net [25]	0.955±0.045	0.876±0.111	**0.927±0.131**	0.889±0.133	0.817±0.150	1.010e−02
UNet3+ [24]	0.957±0.044	0.889±0.149	0.909±0.135	0.893±0.133	0.825±0.150	3.810e−02
TransUNet [39]	0.954±0.047	0.900±0.101	0.906±0.121	0.895±0.099	0.821±0.136	2.886e−02
LeViT-UNet [38]	0.953±0.049	0.889±0.150	0.888±0.147	0.882±0.136	0.808±0.157	1.131e−01
DCSAU-Net	**0.959±0.045**	**0.914±0.098**	0.924±0.077	**0.914±0.077**	**0.850±0.114**	1.000e+00

Table 5
Detailed ablation study of the DCSAU-Net architecture.

Dataset	Method	Accuracy	Precision	Recall	DSC	mIoU	P-values	Parameters	FLOPs	FPS
CVC-ClinicDB	U-Net [22]	0.984±0.019	0.882±0.195	0.893±0.176	0.872±0.189	0.809±0.213	3.182e−02	13.40M	31.11	109.95
	U-Net + PFC	0.987±0.014	0.901±0.191	0.885±0.214	0.881±0.211	0.828±0.216	2.398e−02	13.37M	29.70	103.49
	U-Net + CSA	0.987±0.015	0.890±0.211	0.903±0.179	0.890±0.193	0.840±0.204	1.472e−01	2.62M	8.33	44.26
	U-Net + PFC + CSA (ours)	**0.990±0.015**	**0.917±0.148**	**0.920±0.143**	**0.916±0.141**	**0.861±0.156**	1.000e+00	2.60M	6.91	43.37
SegPC-2021	U-Net [22]	0.939±0.053	0.842±0.142	0.879±0.118	0.855±0.119	0.766±0.148	2.299e−07	13.40M	124.58	48.46
	U-Net + PFC	0.946±0.046	0.866±0.123	0.874±0.086	0.864±0.085	0.780±0.144	1.496e−06	13.37M	119.79	47.63
	U-Net + CSA	0.946±0.046	0.855±0.135	0.896±0.071	0.870±0.080	0.781±0.146	1.152e−04	2.62M	33.35	33.22
	U-Net + PFC + CSA (ours)	**0.950±0.045**	**0.871±0.113**	**0.910±0.067**	**0.886±0.078**	**0.806±0.121**	1.000e+00	2.60M	27.66	32.08
2018 Data Science Bowl	U-Net [22]	0.955±0.047	0.872±0.105	0.920±0.111	0.887±0.090	0.808±0.126	5.039e−05	13.40M	31.11	125.30
	U-Net + PFC	0.955±0.046	0.905±0.105	0.910±0.096	0.901±0.084	0.830±0.123	3.620e−02	13.37M	29.70	117.09
	U-Net + CSA	0.957±0.045	0.903±0.105	**0.925±0.090**	0.908±0.082	0.839±0.122	2.179e−01	2.62M	8.33	43.87
	U-Net + PFC + CSA (ours)	**0.959±0.045**	**0.914±0.098**	0.924±0.077	**0.914±0.077**	**0.850±0.114**	1.000e+00	2.60M	6.91	43.42
ISIC-2018	U-Net [22]	0.952±0.079	0.883±0.152	0.906±0.180	0.874±0.158	0.802±0.182	1.896e−07	13.40M	31.11	115.85
	U-Net + PFC	0.955±0.076	0.915±0.129	0.901±0.148	0.890±0.128	0.821±0.161	6.386e−04	13.37M	29.70	113.36
	U-Net + CSA	0.955±0.078	0.915±0.123	0.909±0.140	0.893±0.127	0.830±0.160	1.068e−02	2.62M	8.33	43.19
	U-Net + PFC + CSA (ours)	**0.960±0.075**	**0.917±0.127**	**0.922±0.139**	**0.904±0.128**	**0.841±0.158**	1.000e+00	2.60M	6.91	41.91
BraTS-2021	U-Net [22]	0.930±0.021	0.752±0.133	0.746±0.129	0.748±0.128	0.650±0.125	4.866e−06	13.40M	27.35	115.38
	U-Net + PFC	0.931±0.025	0.753±0.157	0.765±0.167	0.759±0.149	0.672±0.133	3.799e−03	13.37M	26.10	111.79
	U-Net + CSA	0.933±0.039	**0.793±0.109**	0.764±0.125	0.780±0.117	0.687±0.132	2.535e−02	2.62M	7.33	42.47
	U-Net + PFC + CSA (ours)	**0.935±0.015**	0.792±0.091	**0.785±0.087**	**0.788±0.105**	**0.703±0.092**	1.000e+00	2.60M	6.08	40.75

4.5.4. Comparison on ISIC-2018 dataset

Table 6 shows the quantitative results on ISIC-2018 dataset for the lesion boundary segmentation task. mIoU is an official evaluation metric for the challenge. According to Table 6, DCSAU-Net has an increase of 2.4% over LeViT-UNet in this metric, and 1.8% over UNet3+ in DSC. Within the rest of metrics, our model achieves a recall of 0.922

Table 6
Results on the ISIC 2018 (Skin lesion segmentation challenge).

Method	Accuracy	Precision	Recall	DSC	mIoU	P-values
U-Net [22]	0.952±0.079	0.883±0.152	0.906±0.180	0.874±0.158	0.802±0.182	1.896e−07
Unet++ [23]	0.954±0.077	0.899±0.136	0.906±0.155	0.883±0.138	0.812±0.171	3.861d−05
Attention-UNet [35]	0.954±0.078	0.915±0.140	0.890±0.171	0.883±0.149	0.814±0.180	5.818e−04
ResUNet++ [27]	0.954±0.082	0.905±0.139	0.889±0.183	0.879±0.153	0.810±0.181	1.170e−04
R2U-Net [51]	0.945±0.078	0.834±0.189	0.912±0.163	0.848±0.160	0.762±0.189	2.522e−06
DoubleU-Net [25]	0.953±0.092	0.903±0.149	0.897±0.186	0.879±0.167	0.813±0.191	2.504e−07
UNet3+ [24]	0.956±0.068	0.889±0.151	0.916±0.130	0.886±0.132	0.816±0.165	1.422e−05
TransUNet [39]	0.945±0.085	0.847±0.186	0.898±0.185	0.849±0.178	0.770±0.203	1.450e−08
LeViT-UNet [38]	0.954±0.089	0.896±0.152	0.908±0.176	0.883±0.161	0.817±0.185	8.289e−06
DCSAU-Net	**0.960±0.075**	**0.917±0.127**	**0.922±0.139**	**0.904±0.128**	**0.841±0.158**	1.000e+00

Table 7
Results on the BraTS 2021 (Brain tumour segmentation challenge).

Method	Accuracy	Precision	Recall	DSC	mIoU	P-values
U-Net [22]	0.930±0.021	0.752±0.133	0.746±0.129	0.748±0.128	0.650±0.125	4.866e−06
Unet++ [23]	0.932±0.037	0.765±0.146	0.761±0.135	0.762±0.112	0.671±0.123	2.613e−04
Attention-UNet [35]	0.933±0.028	0.779±0.143	0.751±0.155	0.765±0.178	0.677±0.169	6.295e−04
ResUNet++ [27]	0.934±0.033	**0.802±0.118**	0.746±0.101	0.771±0.114	0.682±0.131	2.174e−02
R2U-Net [51]	0.929±0.082	0.767±0.194	0.756±0.197	0.744±0.214	0.645±0.201	5.449e−07
DoubleU-Net [25]	0.933±0.042	0.771±0.132	0.758±0.096	0.762±0.097	0.675±0.135	3.673e−04
UNet3+ [24]	0.934±0.022	0.743±0.137	0.781±0.146	0.767±0.118	0.679±0.152	7.452e−07
TransUNet [39]	0.933±0.046	0.744±0.126	0.769±0.132	0.758±0.153	0.662±0.146	4.674e−06
LeViT-UNet [38]	0.931±0.034	0.749±0.124	0.754±0.118	0.751±0.111	0.654±0.137	8.263e−07
DCSAU-Net	**0.935±0.015**	0.792±0.091	**0.785±0.087**	**0.788±0.105**	**0.703±0.092**	1.000e+00

Table 8
An investigation of different kernel sizes in the PFC block of the DCSAU-Net architecture.

Dataset	Kernel size	Accuracy	Precision	Recall	DSC	mIoU	P-values	Parameters	FLOPs	FPS
CVC-ClinicDB	3 × 3	0.989±0.014	0.892±0.196	**0.922±0.176**	0.903±0.188	0.857±0.194	1.719e−01	2.58M	6.24	43.02
	5 × 5	0.987±0.010	0.898±0.172	0.916±0.136	0.904±0.159	0.858±0.174	4.581e−01	2.59M	6.50	42.89
	7 × 7	**0.990±0.015**	**0.917±0.148**	0.920±0.143	**0.916±0.141**	**0.861±0.156**	1.000e+00	2.60M	6.91	43.37
	9 × 9	0.988±0.017	0.908±0.160	0.902±0.180	0.894±0.177	0.841±0.198	1.729e−01	2.61M	7.44	43.39
SegPC-2021	3 × 3	0.946±0.058	0.866±0.118	0.882±0.091	0.869±0.075	0.790±0.145	7.519e−05	2.58M	39.42	32.09
	5 × 5	0.948±0.048	0.863±0.122	0.901±0.070	0.877±0.080	0.800±0.131	2.893e−01	2.59M	40.49	32.02
	7 × 7	**0.950±0.045**	**0.871±0.113**	**0.910±0.067**	**0.886±0.078**	**0.806±0.121**	1.000e+00	2.60M	42.10	32.08
	9 × 9	0.946±0.050	0.851±0.134	0.896±0.078	0.868±0.104	0.786±0.153	1.992e−03	2.61M	44.25	31.45
2018 Data Science Bowl	3 × 3	0.958±0.045	0.911±0.101	0.920±0.076	0.911±0.077	0.845±0.115	3.227e−01	2.58M	6.24	43.31
	5 × 5	0.958±0.044	**0.915±0.096**	0.918±0.077	0.912±0.077	0.847±0.114	5.900e−01	2.59M	6.50	43.12
	7 × 7	**0.959±0.045**	0.914±0.098	**0.924±0.077**	**0.914±0.077**	**0.850±0.114**	1.000e+00	2.60M	6.91	43.42
	9 × 9	0.957±0.045	0.908±0.106	0.921±0.083	0.908±0.081	0.841±0.119	4.596e−01	2.61M	7.44	43.08
ISIC-2018	3 × 3	0.958±0.080	0.921±0.112	0.904±0.171	0.893±0.144	0.829±0.173	6.270e−04	2.58M	6.24	42.17
	5 × 5	0.959±0.077	0.919±0.127	0.913±0.149	0.898±0.139	0.836±0.165	1.958e−02	2.59M	6.50	42.12
	7 × 7	**0.960±0.075**	0.917±0.127	**0.922±0.139**	**0.904±0.128**	**0.841±0.158**	1.000e+00	2.60M	6.91	41.91
	9 × 9	0.958±0.080	**0.922±0.117**	0.903±0.164	0.893±0.146	0.830±0.172	4.805e−04	2.61M	7.44	42.63
BraTS-2021	3 × 3	0.933±0.031	0.776±0.107	0.781±0.113	0.779±0.104	0.691±0.119	9.348e−03	2.58M	5.49	41.22
	5 × 5	0.934±0.027	**0.794±0.156**	0.778±0.153	0.783±0.139	0.697±0.142	2.823e−01	2.59M	5.72	40.99
	7 × 7	**0.935±0.015**	0.792±0.091	**0.785±0.087**	**0.788±0.105**	**0.703±0.092**	1.000e+00	2.60M	6.08	40.75
	9 × 9	0.932±0.034	0.772±0.186	0.766±0.191	0.769±0.173	0.682±0.189	6.755e−05	2.61M	6.55	39.12

and an accuracy of 0.960, which is better than other baseline methods. Also, a high recall score is more favourable in clinic applications [52].

4.5.5. Comparison on BraTS-2021 dataset

Brain tumour segmentation is a critical step in MRI analysis. Our DCSAU-Net and other methods are evaluated on BraTS-2021 dataset. A comparison between each model is provided in Table 7. It can be revealed that DCSAU-Net achieves a DSC of 0.788 mIoU of 0.703, which is 1.7% and 2.1% higher than ResUnet++ respectively. Although we use a sub-optimal 2D architecture, there are still significant improvements with the proposed model. Overall, our proposed model demonstrates the highest score in the most of evaluation metrics, including precision and accuracy.

4.6. Ablation study

In this section, we conduct an extensional ablation study on the DCSAU-Net. The number of parameters, floating point operations (FLOPs) and frames per second (FPS) is calculated to investigate the effectiveness of each module in more detail. Table 5 provides the ablation results of five configurations on all five datasets.

4.6.1. Significance of PFC strategy

The PFC Strategy is an essential part of the proposed DCSAU-Net model. It uses residual depthwise separable architecture with a large kernel size to enrich low-level semantic information in the initial down-sampling block and help to generate a more accurate segmentation mask. We compare the network configurations: U-Net and U-Net + PFC to evaluate the efficiency of the PFC strategy. From the mIoU metric in Table 5, PFC shows an improvement of 1.9% on the CVC-ClinicDB dataset, 1.4% improvement on the SegPC-2021, 2.2% improvement on the 2018 Data Science Bowl dataset and 1.9% improvement on the ISIC 2018 dataset. Thus, it can be concluded that the PFC strategy enhances the performance of the original U-Net.

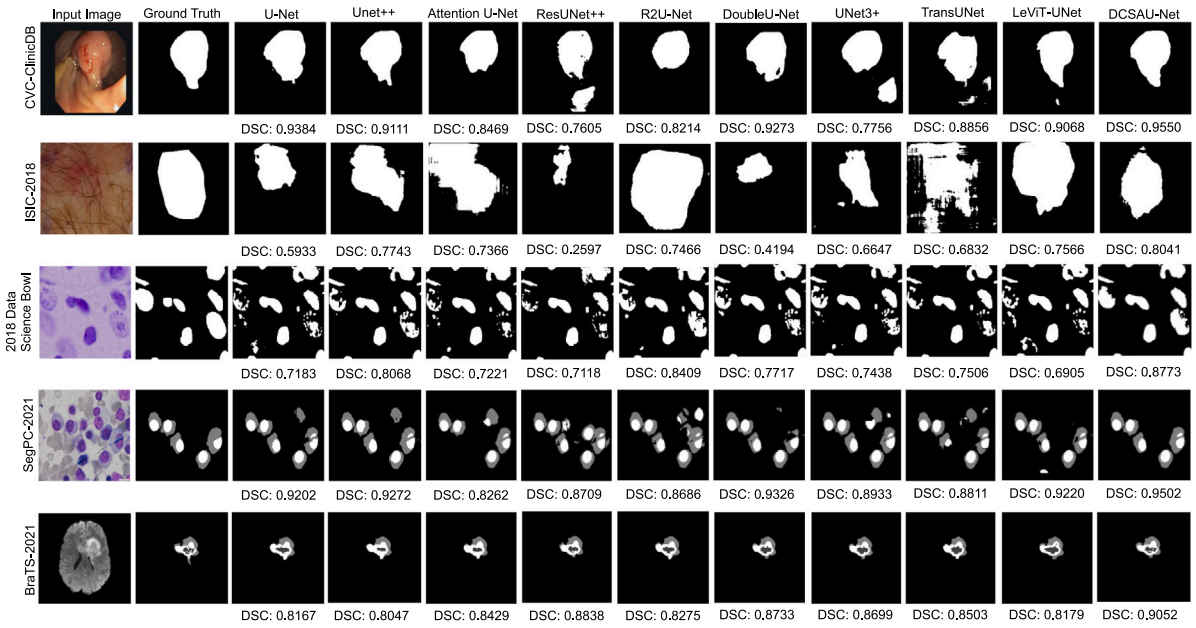

Fig. 4. Qualitative comparison results between DCSAU-Net and other SOTA models on challenging images of five different medical segmentation datasets.

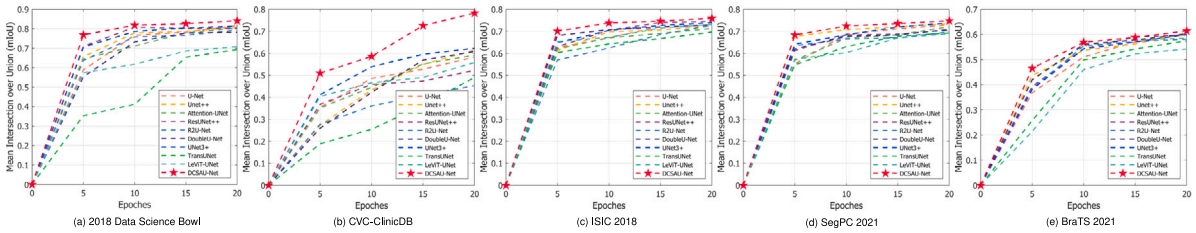

Fig. 5. Visualisations of the first 20 epochs on the test dataset of five medical image segmentation tasks.

4.6.2. *Effectiveness of CSA block*

The DCSAU-Net model uses the CSA block to combine multi-scale feature maps, which can perceive different sizes of lesions in medical images. The effectiveness of CSA block can be evaluated by comparing the configurations: U-Net and U-Net + CSA in Table 5. On the mIoU, the CSA block achieves an improvement of 3.1% on the CVC-ClinicDB dataset, 1.5% improvement on the SegPC-2021, 3.1% improvement on the 2018 Data Science Bowl dataset and 2.8% improvement on the ISIC 2018 dataset. Therefore, we can argue that the CSA block performs better than the U-Net model and has a more significant impact than the PFC strategy. By taking advantage of both modules, the DCSAU-Net model (U-Net + PFC + CSA) can further improve the DSC by 0.6% to 3.5% and the mIoU by 1.1% to 3.3% compared to the U-Net with a single PFC or CSA module.

5. Discussion

Semantic segmentation has been widely witnessed in the field of medical image analysis. Many deep learning models construct encoder–decoder architectures and fuse low-level to high-level semantic information through skip connection. These methods usually select the U-Net [22] block as the header of the encoder to extract low-level semantic information, which probably misses some momentous features in images. Our approach adopts the depthwise separable convolutions with a larger kernel size to build a PFC strategy that retains these primary features as much as possible. In addition, we explore the impact of depthwise convolution with different number of kernel sizes on the performance, which is presented in Table 8. From the experiment results, we can observe that the DCSAU-Net model is able to achieve a similar performance when using 3×3, 5×5 and 7×7 kernel sizes. In practical scenarios, people probably select a small kernel size to reduce the number of parameters and computation costs. However, to display the best performance of our proposed architecture in the study, we use a 7×7 kernel size to train the model. Based on the efficiency of depthwise separable convolution, adding more such layers may improve the information capture capability of the PFC module in the low-level semantic layer, which is worth exploring in future work. We next establish the CSA block that not only enhances the connectivity across different channels but also strengthens the feature representation in different scales with the attention mechanism and completes the multi-scale combination in the end. The effectiveness of both modules has been shown in Table 5 and proved by the ablation study. Although U-Net performs a shorter inference time than the DCSAU-Net model, our approach uses a tiny number of parameters in the equal output feature channels and also expends acceptable inference time, which is more suitable for deployment on machines with limited memory.

To further demonstrate that there is a significant improvement of the DCSAU-Net model for the medical image segmentation task, we visualise some of segmentation results using all models on challenging images, which is provided in Fig. 4. From the qualitative results, the segmentation mask generated by our proposed model is able to capture more proper foreground information from low-quality images, such as incomplete staining or obscurity, compared to other SOTA methods. Although the segmentation result of DCSAU-Net is not completely correct, this imperfect mask with more shape information has the possibility to be fixed using image post-processing algorithms, such as applying conditional random fields. In Fig. 6, we visualise some segmentation results in which DCSAU-Net fails to separate the target

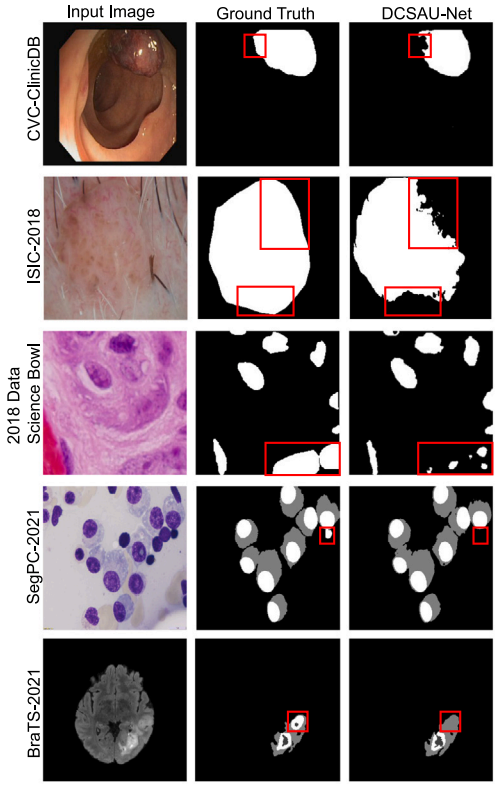

Fig. 6. The scenarios that DCSAU-Net fails to segment the target from images on five medical image segmentation datasets.

from images due to the nuclei with tiny size, or high similarity between the foreground and background. In our experiments, we train all models based on a standard dice loss function. We compare the convergence speed of each model on all five datasets, which is shown in Fig. 5. It can be observed that our proposed model converges noticeably faster than other SOTA methods in the first 20 epochs, which means the DCSAU-Net model is able to reach reliable performance by training fewer epochs. Also, we can prove that compared with other methods, in most cases, the performance improvement of DCSAU-Net is statistical significance with p-value < 0.05. Furthermore, 3D CNNs can efficiently preserve relational information in volumetric medical imaging data. Therefore, extending our framework to 3D is our future work. Also, using other advanced strategies [29] in data pre-processing, model training and inference can further improve the performance in medical image segmentation. In summary, DCSAU-Net shows its robustness and superior performance on various medical segmentation tasks and we believe it can be considered as a new SOTA model for medical image segmentation.

6. Conclusion

In this paper, we propose an encoder–decoder architecture for medical image segmentation, called DCSAU-Net. The presented model is comprised of the PFC strategy and the CSA block. The former enhances the ability to preserve primary features from images. The latter splits the input feature maps into two feature groups. Each group contains a different number of convolutions and highlights meaningful features using the attention mechanism. Therefore, the CSA block can combine feature maps in the different receptive fields. We evaluate our model on five medical image segmentation datasets. The results show that the DCSAU-Net architecture achieves higher scores than other SOTA models in the DSC and mIoU metrics. Especially, our model performs better on the multi-class segmentation task and complex images. In the future, we will focus on optimising the DCSAU-Net architecture for 3D medical image segmentation tasks.

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.

References

[1] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognit. 110 (2021) 107332.
[2] A.S. Coates, E.P. Winer, A. Goldhirsch, R.D. Gelber, M. Gnant, M. Piccart-Gebhart, B. Thürlimann, H.-J. Senn, P. Members, F. André, et al., Tailoring therapies—improving the management of early breast cancer: St Gallen international expert consensus on the primary therapy of early breast cancer 2015, Ann. Oncol. 26 (8) (2015) 1533–1546.
[3] X. Chen, B.M. Williams, S.R. Vallabhaneni, G. Czanner, R. Williams, Y. Zheng, Learning active contour models for medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11632–11640.
[4] S. He, K.T. Minn, L. Solnica-Krezel, M.A. Anastasio, H. Li, Deeply-supervised density regression for automatic cell counting in microscopy images, Med. Image Anal. 68 (2021) 101892.
[5] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, D. Shen, High-resolution encoder–decoder networks for low-contrast medical image segmentation, IEEE Trans. Image Process. 29 (2019) 461–475.
[6] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1) (1979) 62–66.
[7] M. Kass, A. Witkin, D. Terzopoulos, Snakes: Active contour models, Int. J. Comput. Vis. 1 (4) (1988) 321–331.
[8] H.R. Tizhoosh, Image thresholding using type II fuzzy sets, Pattern Recognit. 38 (12) (2005) 2363–2372.
[9] D. Riccio, N. Brancati, M. Frucci, D. Gragnaniello, A new unsupervised approach for segmenting and counting cells in high-throughput microscopy image sets, IEEE J. Biomed. Health Inf. 23 (1) (2018) 437–448.
[10] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, CPFNet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging 39 (10) (2020) 3008–3018.
[11] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[12] S. Chen, G. Bortsova, A. García-Uceda Juárez, G.v. Tulder, M.d. Bruijne, Multi-task attention-based semi-supervised learning for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 457–465.
[13] H. Wang, P. Cao, J. Wang, O.R. Zaiane, Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, (no. 3) 2022, pp. 2441–2449.
[14] J.C. Caicedo, A. Goodman, K.W. Karhohs, B.A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, et al., Nucleus segmentation across imaging experiments: The 2018 data science bowl, Nature Methods 16 (12) (2019) 1247–1253.
[15] N.C. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, et al., Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC), in: 2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018, IEEE, 2018, pp. 168–172.
[16] P. Tschandl, C. Rosendahl, H. Kittler, The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Sci. Data 5 (1) (2018) 1–9.
[17] J. Bernal, F.J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, F. Vilariño, WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians, Comput. Med. Imaging Graph. 43 (2015) 99–111.
[18] A. Gupta, R. Gupta, S. Gehlot, S. Goswami, Segpc-2021: Segmentation of multiple myeloma plasma cells in microscopic images, 2021, http://dx.doi.org/10.21227/7np1-2q42.
[19] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S. Pati, et al., The RSNA-ASNR-Miccai Brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021, arXiv preprint arXiv:2107.02314.
[20] B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., The multimodal brain tumor image segmentation benchmark (BRATS), IEEE Trans. Med. Imaging 34 (10) (2014) 1993–2024.

[21] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J.S. Kirby, J.B. Freymann, K. Farahani, C. Davatzikos, Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features, Sci. Data 4 (1) (2017) 1–13.

[22] 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.

[23] Z. Zhou, M.M. Rahman 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.

[24] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, J. Wu, Unet 3+: A full-scale connected unet for medical image segmentation, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 1055–1059.

[25] D. Jha, M.A. Riegler, D. Johansen, P. Halvorsen, H.D. Johansen, Doubleu-net: A deep convolutional neural network for medical image segmentation, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2020, pp. 558–564.

[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] D. Jha, P.H. Smedsrud, M.A. Riegler, D. Johansen, T. De Lange, P. Halvorsen, H.D. Johansen, Resunet++: An advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia, ISM, IEEE, 2019, pp. 225–2255.

[28] T. Tarasiewicz, M. Kawulok, J. Nalepa, Lightweight u-nets for brain tumor segmentation, in: International MICCAI Brainlesion Workshop, Springer, 2021, pp. 3–14.

[29] F. Isensee, P.F. Jaeger, S.A. Kohl, J. Petersen, K.H. Maier-Hein, NNU-Net: A self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18 (2) (2021) 203–211.

[30] Y. He, D. Yang, H. Roth, C. Zhao, D. Xu, Dints: Differentiable neural network topology search for 3d medical image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5841–5850.

[31] Z. Huang, Z. Wang, Z. Yang, L. Gu, Adwu-Net: Adaptive depth and width U-net for medical image segmentation by differentiable neural architecture search, in: International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 576–589.

[32] 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.

[33] C. Kaul, S. Manandhar, N. Pears, Focusnet: An attention-based fully convolutional network for medical image segmentation, in: 2019 IEEE 16th International Symposium on Biomedical Imaging, ISBI 2019, IEEE, 2019, pp. 455–458.

[34] A. Liu, X. Huang, T. Li, P. Ma, Co-Net: A collaborative region-contour-driven network for fine-to-finer medical image segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 1046–1055.

[35] 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.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[37] Y. Yuan, X. Chen, X. Chen, J. Wang, Segmentation transformer: Object-contextual representations for semantic segmentation, 2019, arXiv preprint arXiv:1909.11065.

[38] G. Xu, X. Wu, X. Zhang, X. He, Levit-UNet: Make faster encoders with transformer for medical image segmentation, 2021, arXiv preprint arXiv:2107.08623.

[39] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A.L. Yuille, Y. Zhou, Transunet: Transformers make strong encoders for medical image segmentation, 2021, arXiv preprint arXiv:2102.04306.

[40] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.

[41] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.

[42] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[43] K. Qi, H. Yang, C. Li, Z. Liu, M. Wang, Q. Liu, S. Wang, X-Net: Brain stroke lesion segmentation based on depthwise separable convolution and long-range dependencies, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 247–255.

[44] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al., MMDetection: Open mmlab detection toolbox and benchmark, 2019, arXiv preprint arXiv:1906.07155.

[45] X. Ding, X. Zhang, J. Han, G. Ding, Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11963–11975.

[46] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[47] Z. Zhang, Q. Liu, Y. Wang, Road extraction by deep residual U-Net, IEEE Geosci. Remote Sens. Lett. 15 (5) (2018) 749–753.

[48] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, P. Torr, Res2net: A new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.

[49] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 2736–2746.

[50] J. Chen, E. Asma, C. Chan, Targeted gradient descent: A novel method for convolutional neural networks fine-tuning and online-learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 25–35.

[51] 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.

[52] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, M. Vallières, S. Zhu, J. Xie, Y. Peng, et al., Head and neck tumor segmentation in PET/CT: The HECKTOR challenge, Med. Image Anal. 77 (2022) 102336.