2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS) | 979-8-3503-1224-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CBMS58004.2023.00237

SACA-UNet:Medical Image Segmentation Network Based on Self-Attention and ASPP

Gaojuan Fan
Henan University
Kaifeng, China
fangaojuan@henu.edu.cn

Jie Wang
Henan University
Kaifeng, China
wangjie@henu.edu.cn

Chongsheng Zhang⋆
Henan University
Kaifeng, China
cszhang@henu.edu.cn

***Abstract*—In recent years, deep learning based techniques have been successfully applied to medical image segmentation, which plays an important role in intelligent lesion analysis and disease diagnosis. At present, the mainstream segmentation models are primarily based on the U-Net model for extracting local features through multi-layer convolution, which lacks global information and the multi-scale semantic information interaction between the Encoder and Decoder process, leading to sub-optimal segmentation performance. To address such issues, in this work we propose a new medical image segmentation network, namely *SACA-UNet*, which improves the U-Net model via the self-attention and cross atrous spatial pyramid pooling (Cross-ASPP) mechanisms. In specific, SACA-UNet first utilizes the self-attention mechanism to capture the global feature, it next devises a Cross-ASPP module to extract and fuse features of varying reception fields to prompt multi-scale semantic interaction. We evaluate the segmentation performance of our proposed model on four benchmark datasets including the *ISIC2018*, *BUSI*, *CVC-ClinicDB*, and *COVID-19* datasets, in terms of both the Dice coefficient and IoU metrics. Experimental results demonstrate that SACA-UNet remarkably outperforms the baseline methods.**

Keywords—Medical image segmentation, Deep Learning, ASPP, Self-Attention mechanism

I. Introduction

Medical image processing and analysis techniques are of great help to medical staff in the pathological analysis and disease diagnosis process [1]. One primary task in medical image processing is image segmentation which aims to automatically locate the lesion or target areas in the medical images so as to assist medical staff in disease diagnosis and the subsequent operations [2]. In recent years, deep learning techniques have been successfully applied to medical image segmentation and yielded significantly better performance than traditional methods.

U-Net [3] and its variants are currently the mainstream approaches to medical image segmentation. U-Net adopts the symmetrical U-shaped structure in the encoding and decoding processes, it preserves high-level semantic information through skip connections. However, the convolution and pooling layers may cause the loss of global spatial information. To solve this problem, U-Net++ [4], and UNeXt [5] have been proposed, yet the global information loss issue still remains. Researchers have also investigated the combination of CNN and self-attention for capturing the global features of the images [6]–[8]. The ASPP (atrous spatial pyramid pooling) [9], [10] mechanism can effectively enlarge the receptive fields of the network by using different dilation rates, which can capture the global information of the images and improves the segmentation performance in turn.

In this work, we propose a new medical image segmentation network, namely SACA-UNet, which improves upon U-Net via the self-attention and Cross-ASPP mechanisms to tackle the problems of lack of global features and multi-scale semantic information interactions. SACA-UNet devises a global feature learning module, namely Global-Block, and a Cross-ASPP module to better capture the global context information and strengthen the connections of the multi-scale semantic features at different reception fields.

Finally, we carry out extensive experiments and ablation studies on four benchmark datasets, which demonstrate the outstanding performance of our proposed method in medical image segmentation.

II. Related Work

In this section, we introduce the related work of this work, including deep learning based medical image segmentation, self-attention mechanism and ASPP.

A. Deep Learning based Medical Image Segmentation

Accurate medical image segmentation is of great significance for disease diagnosis and lesion analysis. The early medical image segmentation methods are mainly contour-based algorithms [11] or utilize traditional machine learning algorithms [13]. In the last decade, deep learning has reshaped image processing techniques, they have also demonstrated outstanding performance in medical image segmentation. In particular, Ronneberger et al. [3] proposed the U-Net model for medical image segmentation, which is composed of an encoder-decoder network with skip connections structure and has been proved to be effective for many different segmentation tasks. However, when dealing with small targets, the network performance is lower and there is a lack of attention to spatial information. Zhou et al. [4] proposed U-Net++ model, which redesigned the dense skip connections structure, aiming at reducing the semantic gap between encoder and decoder and improving the accuracy. Valanarasu et al.proposed the UNeXt

Corresponding author.

2372-9198/23/$31.00 ©2023 IEEE
DOI 10.1109/CBMS58004.2023.00065

[5] model, using a combination of MLP and a small amount of convolution to construct a simple and fast medical image segmentation method.

At present, the medical image segmentation method based on deep learning has achieved great success in the field of medical image segmentation due to its powerful representation ability. However, they also face a common problem, due to the inherent limitations of convolution operations, they suffer from the information simplification problem and lacks global information/features.

B. Self-Attention Mechanism

Due to the versatility and efficiency of deep learning, its segmentation results are superior to traditional segmentation algorithms and have been applied to many fields such as medical segmentation. Vaswani et al.proposed the Transformer [14] method and made a breakthrough in the field of NLP. In order to apply Transformer to computer vision tasks, Dosovitskiy et al.proposed the Vision Transformer (ViT) [15] model, which directly applies global Self-Attention to images, adds position coding information, calculates through multi-head attention, and implements the most advanced classification task. Compared with deep learning-based methods, ViT has the problem of pre-training on its own massive datasets. Researchers have tried to introduce Self-Attention into medical image segmentation [7], [8] to improve the accuracy of network segmentation. Chen et al.proposed the TransUNet [16] model, which combines Transformer with CNN as encoder for medical image segmentation.

Through the self-attention mechanism, the problem of excessive localization of encoder and lack of global semantic features in medical image segmentation is alleviated to a certain extent. However, in medical image segmentation, the encoding network still has the problem of semantic information being too independent in the association of deep features.

C. Atrous Spatial Pyramid Pooling

Chen et al. investigated the semantic segmentation task in DeepLab [9], [10], in which they propose ASPP (Atrous Spatial Pyramid Pooling) to robustly segment objects of multiple scales. It uses multiple parallel dilated convolution layers with different sampling rates. The difference between dilated convolution and traditional convolution is that the receptive field of the former is larger and can thus capture more global feature information. Moreover, strengthening the semantic correlations between features in the different layers is also beneficial to image segmentation.

III. Methodology

The overall architecture of our proposed SACA-UNet network is shown in Fig. 1. First, in the encoder of our network, each image will first pass through three convolution modules. Each convolution module contains Conv3×3, normalization, pooling, and ReLU activation function to extract the local features from the image. Next, with the designed double Global-Block modules in the encoder, we focus on the most important features using the attention mechanism and obtain the global feature. We then devise the Cross-ASPP module to progressively fuse features of different receptive fields and obtain the global feature, which can combine the multi-scale semantics and strengthen the feature representation capability. It is also beneficial for the segmentation of small objects.

In the decoder of our network, we first adopt the double Global-Block modules to find the most important features from the ones obtained through Cross-ASPP, next apply upsampling to gradually recover the feature dimension. To this end, each module uses the bilinear interpolation method to perform upsampling operation, it also combines the upsampled features with the intermediate features of the same dimension obtained in the encoder stage.

In the following, we will illustrate the SACA-UNet structure in four parts: the convolution module, the Global-Block module, the skip connections and the Cross-ASPP module.

A. The Convolution Stage

The SACA-UNet network passes the original segmented image through three convolution modules, each of which includes a convolution layer, a batch normalization layer, pooling and a ReLU activation function. At the encoding stage, the convolution module will gradually shrink the feature dimension, meanwhile increase the channel number. Such convolution operations can extract features of different dimensions at different layers. At the decoding stage, a bilinear interpolation layer is used upon each module for feature upsampling, to gradually recover the feature dimension.

B. The Global-Block Module

The feature map obtained by the convolution modules in the encoder will be sent to the Global-Block module to find the most important global features, as shown in Fig. 2.

The input feature shape of the module (B, C, H, W), which B represents the number of feature maps in a batch (BatchSize), C represents the number of channels of a feature map, H and W represents the height and width of the input feature map, respectively. The input feature map X will be processed by the double Global-Block modules to extract the global feature maps Z via the Self-Attention mechanism (Token Self-Attn). After channel attention (Channel Attn), the feature Z is residually connected with the input feature X (added by elements) and obtains Z'. After MLP, Z' is residually connected with itself again (added by elements) to obtain feature G.This module uses two attention mechanisms to focus on the context information of global features. Where $Q, K, V \in R^d$ represents query matrix, key matrix and value matrix, respectively. d represents the dimension of the query or key, and the value in B is taken from the offset matrix. The Global-Block module can be formulated as:

$$Z = MatMul(SoftMax(\frac{Q^T K}{\sqrt{d}})V, linear(x)) \tag{1}$$

$$Z' = SoftMax(\frac{Q^T K}{\sqrt{d}})V + x \tag{2}$$

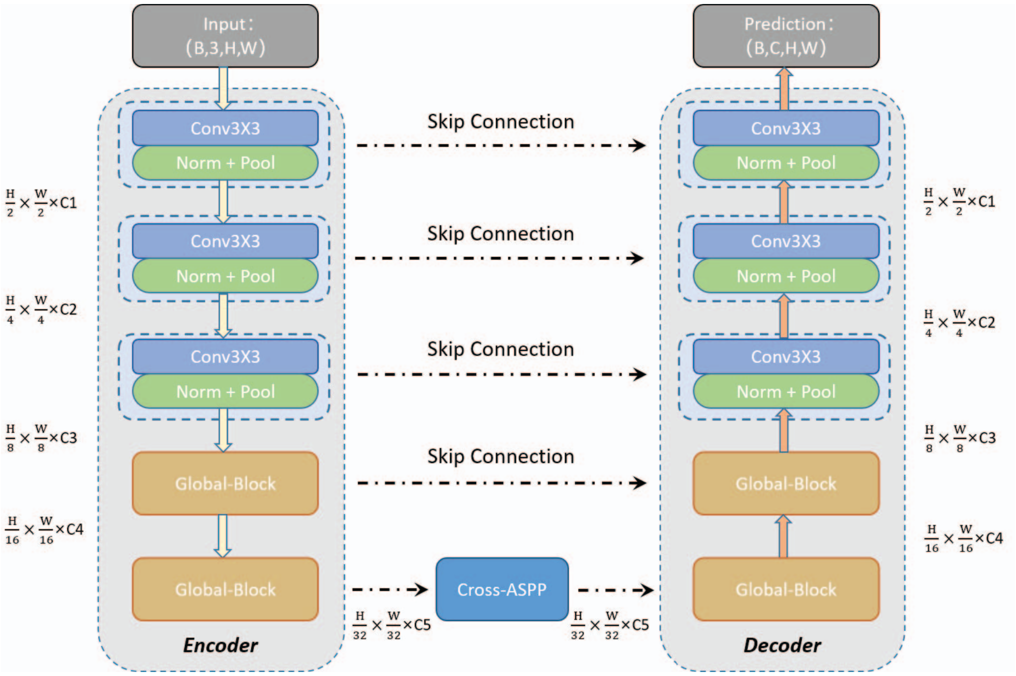

Input：
(B,3,H,W)
Conv3X3
Norm + Pool
Conv3X3
Norm + Pool
Conv3X3
Norm + Pool
Global-Block
Global-Block
Encoder
$\frac{H}{2}\times\frac{W}{2}\times C1$
$\frac{H}{4}\times\frac{W}{4}\times C2$
$\frac{H}{8}\times\frac{W}{8}\times C3$
$\frac{H}{16}\times\frac{W}{16}\times C4$
Skip Connection
Skip Connection
Skip Connection
Skip Connection
Cross-ASPP
$\frac{H}{32}\times\frac{W}{32}\times C5$
$\frac{H}{32}\times\frac{W}{32}\times C5$
Prediction：
(B,C,H,W)
Conv3X3
Norm + Pool
Conv3X3
Norm + Pool
Conv3X3
Norm + Pool
Global-Block
Global-Block
Decoder
$\frac{H}{2}\times\frac{W}{2}\times C1$
$\frac{H}{4}\times\frac{W}{4}\times C2$
$\frac{H}{8}\times\frac{W}{8}\times C3$
$\frac{H}{16}\times\frac{W}{16}\times C4$

Fig. 1. The SACA-UNet Framework.

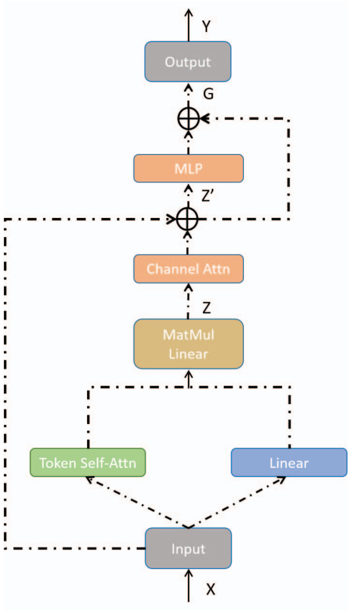


Fig. 2. Global-Block Module structure.

$$G = MLP(Z') + Z' \tag{3}$$

Among them, (1) represents feature X passes through the Self-Attention(Token Self-Attn) block and the Linear block respectively, and then passes through the output of matrix multiplication. (2) represents the output of feature X and (1) element-wise addition (residual connection) through Channel Attn.(3) represents (2) element-wise addition through MLP and (2). The calculation of Self-Attention is as follows:

$$Attention(Q,K,V) = SoftMax(\frac{Q^T K}{\sqrt{d}} + B)V \tag{4}$$

C. Skip Connection

The skip connections structure is used to fuse multi-scale features in the encoder and the upsampled features from the corresponding decoding module to reduce the loss of spatial information caused by upsampling.

D. Cross-ASPP Module

The Cross-ASPP module, as shown in Fig. 3, connects the encoder and the decoder. It is mainly composed of three heads, which are two Conv1×1(left and right heads) heads, and one three dilated convolutions (middle) head. The two Conv1×1 head can preserve the basic features, while three dilated convolutions head uses receptive fields of different scales to extract and fuse multi-scale features, leading to stronger feature representations. Finally, we fuse the features from the three heads and adopt another Conv1×1 to obtain more representative features of the image. In order to ensure that the size of the feature map obtained by atrous convolution with different dilation rates is the same, reset to padding equals dilation.

In the ASPP operation, we gradually fuse feature maps of neighboring scales residually, instead of concatenating the feature maps of all the scales, which is referred to as Cross-ASPP. It can further reduce the loss of feature information.

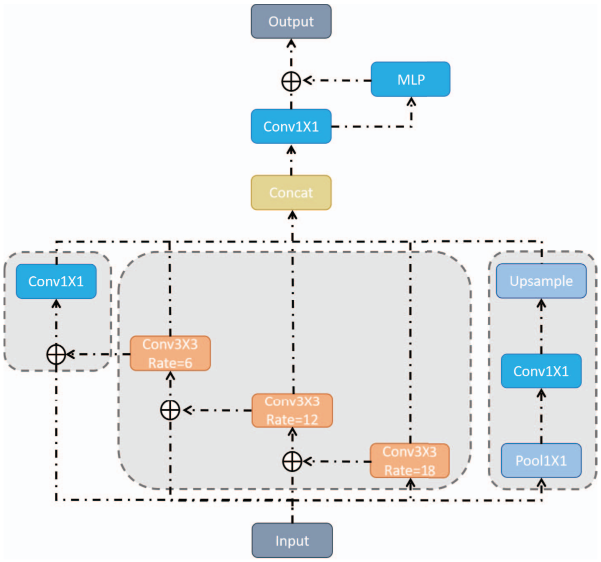


Fig. 3. The Cross-ASPP Module.

Finally, we will obtain strong and representative feature representations, which will be sent to the decoder block for capture the attention and recover the feature dimension.

E. Loss Function

We use the combination of the Binary Cross Entropy Loss (BCE Loss) [20] and the Dice Loss [21] to train our segmentation model. Let $\hat{y}$ be the predicted label and y be the ground-truth label, L be the overall loss.

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i log\left(p_i\right) + \left(1 - y_i\right) log(1 - p_i)\right] \quad (5)$$

$$DiceLoss = 1 - Dice = 1 - \frac{2\left|y \cap \hat{y}\right|}{|y| + |\hat{y}|} \quad (6)$$

$$L = 0.5BCE(\hat{y}, y) + DiceLoss(\hat{y}, y) \quad (7)$$

IV. Experiments and Results

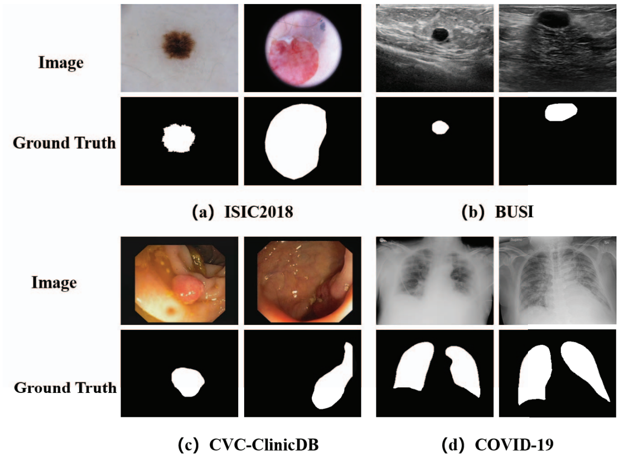
All the experiments are conducted on an NVIDIA GeForce RTX 3070 GPU work station, running Linux Mint OS. We implement our network based on Python 3.7.0 and Pytorch 1.9.0. For experimental setup, we use the Adam optimizer, and the initial lr=0.0001, batchsize=8, momentum=0.9, epoch=200. Each datasets is averaged by three experiments.

A. Datasets and Evaluation Metrics

In order to verify the effectiveness of SACA-UNet, four publicly available medical image datasets are used in our experiments, as shown in Fig. 4. ISIC2018 [17] is a skin lesion datasets, BUSI [18] is a breast ultrasound image datasets, including ultrasound images of normal, benign and malignant pathology of breast cancer and corresponding segmentation maps. The CVC-ClinicDB [19] datasets extracts polyp images from colonoscopy detection videos. COVID-19 novel coronavirus pneumonia datasets. The specific information is shown in Table I.

TABLE I
Dataset Information

Dataset	Amount	Size	Format
ISIC2018	2594	512×512	.jpg
BUSI	647	256×256	.png
CVC-ClinicDB	612	256×256	.png
COVID-19	3616	256×256	.png



(a) ISIC2018 (b) BUSI (c) CVC-ClinicDB (d) COVID-19

Fig. 4. Datasets.

The evaluation metrics used in the experiments are the Dice coefficient and the IoU metric. Both metrics are adopted to measure the segmentation performance of different methods.

$$Dice = \frac{2\left|A \cap B\right|}{|A| + |B|} = \frac{2TP}{2TP + FP + FN} \quad (8)$$

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \quad (9)$$

Among them, A and B are the number of pixels in the predicted result and the true result respectively; TP is the number of true positive samples; FP and FN were the number of false positive samples and false negative samples, respectively

B. Ablation Experiment

In order to compare the influence of different modules on the performance of our model, we conduct ablation experiments on four datasets. The results are shown in Table II and Table III.

We first show the advantages of Cross-ASPP with respect to ASPP. In Table II, we see that, compared with ASPP, our designed Cross-ASPP module improves the segmentation performance by 0.5% and 0.8% on the ISIC2018 and BUSI

TABLE II
ABLATION EXPERIMENTS OF DIFFERENT STRUCTURES.

Method	ISIC2018		BUSI		CVC-ClinicDB		COVID-19	
	Dice	IoU	Dice	IoU	Dice	IoU	Dice	IoU
UNeXt	0.8970	0.8170	0.7937	0.6695	0.9006	0.8201	0.9739	0.9497
(Global-Block)	0.8887	0.8015	0.8150	0.6947	0.9103	0.8394	0.9743	0.9550
(Global-Block+ASPP)	0.8960	0.8130	0.8339	0.7188	**0.9226**	**0.8579**	0.9757	0.9529
SACA-UNet	**0.9017**	**0.8219**	**0.8424**	**0.7314**	0.9207	0.8535	**0.9760**	**0.9536**

datasets, respectively, in terms of the Dice coefficient. Similar observations also hold on the IoU metric. Moreover, the combination of Cross-ASPP and Global-Block can obtain an overall segmentation performance increase of 0.47%, 4.97%, 2.01%, 0.21% on the ISIC2018, BUSI, CVC-ClinicDB and COVID-19 datasets, respectively, in terms of the Dice metric.

TABLE III
ABLATION EXPERIMENTS FOR THE NUMBER OF SKIP CONNECTIONS.

Skip connection	BUSI		CVC-ClinicDB	
	Dice	IoU	Dice	IoU
0	0.8122	0.6886	0.9014	0.8217
1	0.8216	0.7022	0.9041	0.8260
2	0.8210	0.7021	0.9078	0.8345
3	0.8276	0.7125	0.9117	0.8388
4	**0.8424**	**0.7314**	**0.9207**	**0.8535**

From Table III, we can see the influence of the number of skip connections on the segmentation accuracy of the SACA-UNet network. The segmentation performance of SACA-UNet network increases with the number of skip connections. Therefore, in order to make the model more robust, the number of skip connections is set to 4 in this paper.

C. Main Results

In order to verify the effectiveness of the SACA-UNet network on four different datasets, SACA-UNet is compared with U-Net, Attention UNet [22], UNeXt and MTUNet [23], as shown in Table IV.

As can be seen from table IV, the Dice value of SACA-UNet network is 6.14% higher than that of U-Net on ISIC2018, and 7.89 % higher on BUSI, and 4.26 % higher on CVC-ClinicDB. The IoU value is 7.64 % higher than that of U-Net on ISIC2018, 9.29 % higher on BUSI, and 6.54 % higher on CVC-ClinicDB. On the COVID-19 dataset, SACA-UNet uses 1/4 of the epochs needed by U-Net, with a performance drop of 0.27 % in terms of Dice and 0.52 % in terms of IoU. The experimental results show that the SACA-UNet network can reduce the global semantic feature loss.

In Fig. 5, we show the segmentation effects of SACA-UNet network, U-Net network and UNeXt network on the four datasets, respectively. It can be seen that U-Net and UNeXt do not fully retain the edge details in the ISIC2018 skin lesions, and there is more noise information, and the skin lesion area is not accurately segmented. The proposed SACA-UNet network can better maintain the integrity of the target with large changes in size and shape of skin lesions, and the segmentation results are more accurate. U-Net does not retain the edge information of BUSI breast ultrasound completely, and there is a lack of segmentation area. Both our SACA-UNet network and UNeXt network can maintain the integrity of black and white ultrasonic image targets, but our segmentation results are more accurate.

TABLE IV
SEGMENTATION ACCURACY OF DIFFERENT METHODS.

Method	ISIC2018		BUSI	
	Dice	IoU	Dice	IoU
U-Net	0.8403	0.7455	0.7635	0.6385
Attention UNet	0.8899	0.8032	0.8175	0.6993
UNeXt	0.8970	0.8170	0.7937	0.6695
MTUNet	0.8984	0.8166	0.8066	0.6827
SACA-UNet	**0.9017**	**0.8219**	**0.8424**	**0.7314**
Method	**CVC-ClinicDB**		**COVID-19**	
	Dice	IoU	Dice	IoU
U-Net	0.8781	0.7881	**0.9787**	**0.9588**
Attention UNet	0.9112	0.8399	0.9773	0.9563
UNeXt	0.9006	0.8201	0.9739	0.9497
MTUNet	0.9096	0.8385	0.9785	0.9585
SACA-UNet	**0.9207**	**0.8535**	0.9760	0.9536

Overall, comparing with U-Net and UNeXt, SACA-UNet can clearly segment the target area on the four different datasets. It has better segmentation effect on the lesion area, and the boundary area of the lesion is segmented and refined, which is more consistent with the Ground Truth.

V. CONCLUSION

U-Net based segmentation models suffer from the lack of global feature information and low interactions among the

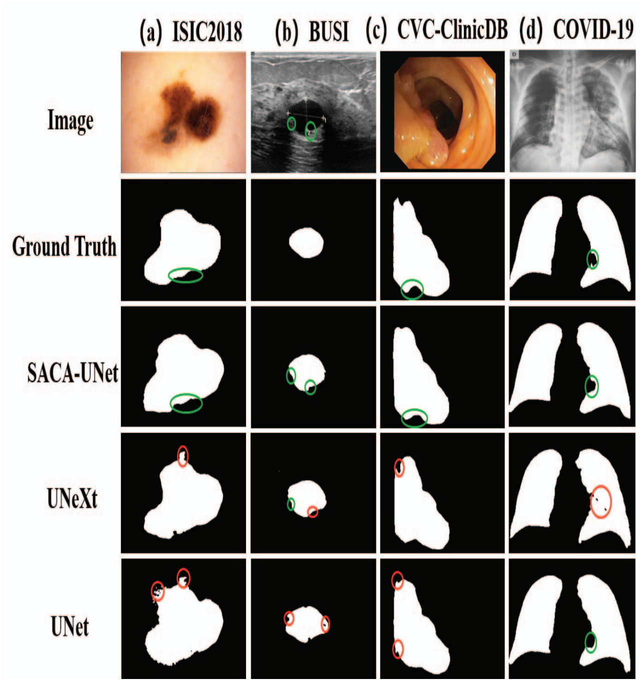


Fig. 5. Segmentation effects of SACA-UNet, UNeXt and U-Net on different datasets.

features of different reception fields. To tackle these problems, in this work we propose the SACA-UNet network that uses the self-attention mechanism and Cross-ASPP for more accurate medical image segmentation. The Global-Block module extracts global context feature map from the input image, while the Cross-ASPP module uses multi-scale information interaction to enhance the semantic correlations between features at different reception fields. Experiments on four benchmark datasets show that our method not only ensures the integrity of the segmented region, but also achieves better overall segmentation performance. Future work will directed towards adapting our proposal to 3D medical data segmentation.

REFERENCES

[1] DALCA A V, GUTTAG J, and SABUNCU M R. Anatomical priors in convolutional networks for unsupervised biomedical segmentation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake, USA, 9290–9299. doi: 10.1109/CVPR.2018.00968. 2018.

[2] LE M, UNKELBACH J, AYACHE N GPSSI: gaussian process for sampling segmentations of images[C]//18th International Conference on Medical Image Computing and Conputer-Assisted Intervention. Munich:Springer, 38-46, 2015.

[3] O. RONNEBERGER, P. FISCHER, T. BROX. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 234-241,2015.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang . U-Net++: a nested U-Net architecture for medical image segmentation.2018.

[5] J. M. J. Valanarasu, V.M. Patel Unext: Mlp-based rapid medical image segmentation network[C]//Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. Cham: Springer Nature Switzerland, 23-33, 2022.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc, 2017.

[7] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and Patel, V. M. Medical transformer: Gated axial-attention for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer International Publishing. pp. 36-46, 2021.

[8] Y. Zhang, H. Liu, and Q. Hu. Transfuse: Fusing transformers and cnns for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings. Springer International Publishing.Part I 24 ,pp. 14-24, 2021.

[9] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4), 834-848, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916, 2015.

[11] Zhang, Tao, Shiqing Wei, and Shunping Ji. ”E2ec: An end-to-end contour-based method for high-quality high-speed instance segmentation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[12] Zichen Liu, Jun Hao Liew, Xiangyu Chen, and Jiashi Feng. Dance: A deep attentive contour model for efficient instance segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 345–354, 2021.

[13] Sida Peng, Wen Jiang, Huaijin Pi, Xiuli Li, Hujun Bao and Xiaowei Zhou. Deep snake for real-time instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8533–8542, 2020.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc, 2017.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T.Unterthiner, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, and Y. Zhou. TransUNet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[17] 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 biomed ical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE 2018.

[18] W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy: Dataset of breast ultrasound images. Datas in brief 28, 104863, 2020.

[19] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43, 99-111, 2015.

[20] Mannor S, Peleg D, Rubinstein R. The cross entropy method for classification[C] //Proceedings of the 22nd International Conference on Machine Learning. New York: ACM Press, 561-568,2005.

[21] Milletari F, Navab N, Ahmadi S A. V-Net: fully convolutional neural networks for volumetric medical image segmentation[C] //Proceedings of the 4th International Conference on 3D Vision. Los Alamitos: IEEE Computer Society Press, 565-571,2016.

[22] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K.,and Rueckert, D. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999. 2018.

[23] Wang, H., Xie, S., Lin, L., Iwamoto, Y., Han, X. H., Chen, Y. W., and Tong, R. Mixed transformer u-net for medical image segmentation. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2390-2394).2022.