Computers in Biology and Medicine 167 (2023) 107596

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

SUnet: A multi-organ segmentation network based on multiple attention

Xiaosen Li [a,b], Xiao Qin [c], Chengliang Huang [d], Yuer Lu [b], Jinyan Cheng [b], Liansheng Wang [e], Ou Liu [b], Jianwei Shuai [b,**], Chang-an Yuan [c,f,*]

[a] School of Artificial Intelligence, Guangxi Minzu University, Nanning, 530006, China
[b] Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou, 325105, China
[c] Guangxi Key Lab of Human-machine Interaction and Intelligent Decision, Nanning Normal University, Nanning, 530023, China
[d] Academy of Artificial Intelligence, Zhejiang Dongfang Polytechnic, Wenzhou, 325025, China
[e] Department of Computer Science, Xiamen University, Xiamen, 361005, China
[f] Guangxi Academy of Science, Nanning, 530007, China

ARTICLE INFO

Keywords:
Medical image segmentation
Transformer
Attention mechanism
Network architecture
Computed tomography

ABSTRACT

Organ segmentation in abdominal or thoracic computed tomography (CT) images plays a crucial role in medical diagnosis as it enables doctors to locate and evaluate organ abnormalities quickly, thereby guiding surgical planning, and aiding treatment decision-making. This paper proposes a novel and efficient medical image segmentation method called SUnet for multi-organ segmentation in the abdomen and thorax. SUnet is a fully attention-based neural network. Firstly, an efficient spatial reduction attention (ESRA) module is introduced not only to extract image features better, but also to reduce overall model parameters, and to alleviate overfitting. Secondly, SUnet's multiple attention-based feature fusion module enables effective cross-scale feature integration. Additionally, an enhanced attention gate (EAG) module is considered by using grouped convolution and residual connections, providing richer semantic features. We evaluate the performance of the proposed model on synapse multiple organ segmentation dataset and automated cardiac diagnostic challenge dataset. SUnet achieves an average Dice of 84.29% and 92.25% on these two datasets, respectively, outperforming other models of similar complexity and size, and achieving state-of-the-art results.

1. Introduction

The abdomen and thorax are regions that contain a majority of human organs and are also prone to various diseases. Segmenting organs from computed tomography (CT) scans plays a crucial role in diagnosis and treatment. However, it is a laborious and error-prone task for doctors [1]. Therefore, there is an urgent need for automated organ segmentation methods in clinical practice to assist doctors in more efficient and accurate diagnosis. In recent years, although artificial intelligence has achieved promising results in biomedicine, such as single-cell multi-omics data analysis [2,3], RNA-RNA interaction [4–6], proteomics research [7–10], biomarker discovery [11], gene/protein signaling network [12,13] and pharmacometabolomics data processing [14], automatic segmentation of abdominal or thoracic organs remains a challenging task. Numerous factors contribute to these challenges, including interference from surrounding tissues, organ deformation or displacement, and low image contrast leading to unclear boundaries. These challenges pose difficulties in achieving accurate and robust organ segmentation.

Previously, several two-dimensional (2D) medical image segmentation models based on convolutional neural networks (CNN) have been proposed, among which the Unet model stands as the most representative one [15]. The Unet incorporates a distinctive U-shaped encoder-decoder structure and skip connections, resulting in improved performance of the model while simultaneously introducing a new design approach for medical image processing. The outstanding performance of the Unet, has subsequently inspired the development of several variant networks. For example, Unet++ enhances model performance by using nested and dense skip connections in place of the original ones [16]. Unet3+ is a U-shaped medical image segmentation model that adopts full-scale skip connections and depth supervisions [17]. In addition, ResUnet [18] and ResUnet++ [19] are excellent medical image segmentation models also based on Unet. However, despite their achievement, these CNN-based models encounter

* Corresponding author. Guangxi Key Lab of Human-machine Interaction and Intelligent Decision, Nanning Normal University, Nanning, 530023, China.
** Corresponding author.
E-mail addresses: shuaijw@wiucas.ac.cn (J. Shuai), 68852917@qq.com (C.-a. Yuan).

https://doi.org/10.1016/j.compbiomed.2023.107596
Received 20 July 2023; Received in revised form 13 September 2023; Accepted 17 October 2023
Available online 18 October 2023

limitations in effectively establishing long-distance dependencies due to the inherent constrains of convolutional operations, which restrict their performance.

Hence, the Transformer [20], known for its proficient ability to capture long-distance dependencies effectively, has garnered attention from researchers in computer vision. The vision transformer (ViT) [21], as the first deep learning model to incorporate transformer in image processing, partitions the image into patches and employs self-attention for feature extraction. The transformer in ViT, which has demonstrated remarkable performance, provided researchers with a fresh approach to feature extraction that overcomes the inherent constraints of convolution. The TransUnet [22], pioneering the integration of transformer into medical image segmentation, substitutes the encoder of the Unet model with a CNN-Transformer hybrid model. This hybrid feature extraction network allows for the extraction of more semantic features, consequently enhancing the model's performance. Then, the Swin-Unet model proposed by Cao et al. [23], a fully transformer-based encoder-decoder architecture for medical image segmentation, has achieved substantial performance. Similarly, several transformer-based medical image segmentation models have been applied to various modalities of medical data [24–31]. Nevertheless, these transformer-based models often have more significant parameter and computational requirements than CNN-based medical image segmentation models, posing a formidable challenge for practitioners with limited computing resources.

Inspired by the pyramid vision transformer (PVT) [32], we have devised a novel encoder-decoder medical image segmentation network referred SUnet. This network uses efficient spatial reduction attention (ESRA), leading to improve performance and reduce model parameters. Furthermore, we propose an efficient multi-attention feature fusion module that effectively merges low-level semantic features from skip connections with high-level semantic features from the decoder's upsampling layers. The experimental results indicate that SUnet outperforms to previous 2D medical imaging models, achieving the best results with equivalent model parameters. The main contributions of this paper are outlined as follows:

1) We propose SUnet, a pure transformer-based U-shaped medical image segmentation network incorporating efficient spatial reduction attention and multi-attention feature fusion.
2) We present efficient spatial reduction attention, which allows the model to perform better while maintaining fewer parameters, and alleviates the overfitting commonly observed in transformer-based models.
3) To reduce computational complexity and data dependence, and to extract more task-related features, we provide an enhanced attention gate (EAG) module based on grouped convolution and residual connections.
4) We propose an efficient feature fusion (EFF) module based on multi-attention, which achieves better fusion between skip connections and decoder features in U-shaped networks.

2. Related work

2.1. Unet

Unet is a deep learning model based convolutional neural network proposed by Ronneberger et al. in 2015 [15]. It is widely used for medical image segmentation tasks, such as abdominal multi-organ segmentation, automatic heart diagnosis, retinal vascular segmentation and skin cancer segmentation. The Unet model is characterized by the encoder-decoder structure and skip connections, which transmit multi-level semantic features of the encoder to the decoder. This

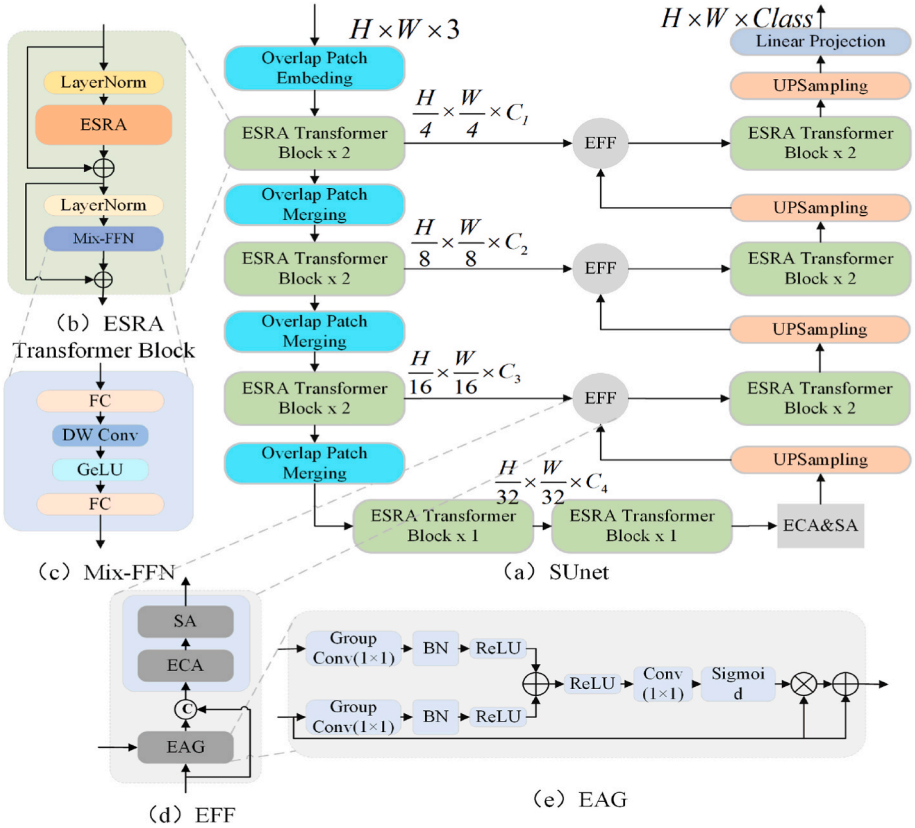

Fig. 1. Overview of SUnet Architecture. (a) is a detailed frame diagram of SUnet medical image segmentation model, which mainly consists of encoder, decoder and EFF module. (b) is a transformer block composed of Spatial reduction attention and Mix-FFN. (c) is the internal details of the Mix-FFN structure. (d) is the composition diagram of the EFF module. (e) is the internal structure of EAG.

operation enables the integration of more low-level semantic features into the feature map, ensuring high accuracy in medical image segmentation. Because of its strong performance in image segmentation, it is also widely used in satellite image segmentation [33,34] and industrial defect detection [35]. Over time, numerous deep learning models have evolved from Unet [36–41]. The SUnet medical image segmentation model proposed in this paper also inherits the U-shaped structure and skip connections of the Unet.

2.2. Attention mechanism

The attention mechanism is an algorithm widely used in deep learning, that mimics the human attention mechanism. In deep learning, all input features are treated equally, regardless of their relevance to the current task. Consequently, the model fails to allocate attention effectively to the task-related areas, which limits the performance of the deep learning model to a certain extent. In contrast, the attention mechanism assigns varying degrees of attention and weighting to different parts of the input, paying more attention to the information related to the current task. The spatial attention (SA) mechanism is a widely used technique in the field of deep learning [42]. In 2018, Oktay et al. proposed the attention gate (AG), which can focus on the channel information in computer vision task [36]. It uses a learning weight coefficient to weight the product of the original input and the selected vector, achieving channel selection and weighting. The squeeze and excitation neural network (SENet) proposed by Hu et al. mainly learns the correlation between channels in the convolutional neural network, and allocates larger weights to channels that are useful for the current task [43]. It is achieved in two steps: squeeze and excitation. Wang et al. proposed an efficient channel attention (ECA) module [44]. The ECA network establishes local contextual correlations among channels by applying a one-dimensional convolution operation to each channel's feature map, thereby achieving adaptive calculation of channel attention. In addition to the attention mechanisms that focus solely on a specific dimension mentioned above, multi-dimensional attention mechanism has also been explored. For example, Wu et al. proposed a multi-dimensional mixed attention convolutional block attention module (CBAM) with a focus on channel and spatial information [45]. Unlike the SENet attention mechanism that exclusively considers channel dimension, CBAM consolidates multiple dimensions of information to better focus on useful information. Also, for the first time, Rahman et al. proposed a hierarchical cascaded attention-based decoder for the first time [46], which also provides ideas for the design of EFF module.

2.3. Pyramid vision transformer

The PVT is a backbone network proposed by Wang et al. [32]. The core module of PVT involves feature compression of the key and value in the multi-head self-attention (MHSA) mechanism via convolutions. This compression operation procedure significantly diminishes both the parameter and computational complexity. The reduction ratio determines the size of the convolutional kernels and strides. The PVT is widely employed as a backbone network in various visual tasks, such as object detection and localization, remote sensing image classification, and medical image analysis. The widespread adoption of PVT underscores its versatility and effectiveness, rendering it a research topic of considerable significance within the field of computer vision.

3. Methods

In this section, we present the model design of the proposed SUnet, with particular attention given to the design of the ESRA transformer block and EFF module. Within the EFF module, we primarily introduce the proposed EAG module.

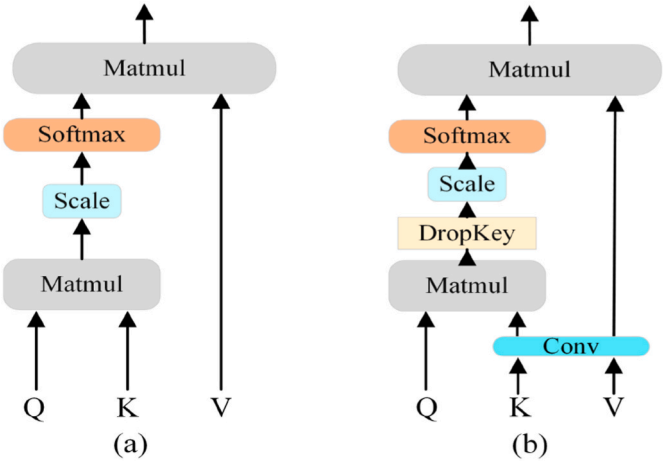

Fig. 2. The structural comparison between self-attention and ESRA is illustrated in the following figures. Figure (a) represents the schematic diagram of self-attention, while figure (b) represents the schematic diagram of ESRA with spatial compression and dropkey.

3.1. Overall architecture

The overall architecture of the proposed SUnet model is shown in Fig. 1(a). SUnet follows the encoder-decoder structure of Unet and employs skip connections to convey low-level semantic information. The channel numbers, denoted as C_1, C_2, C_3, and C_4 are defined as $C_1 = 64$, $C_2 = 128$, $C_3 = 320$, $C_4 = 512$. The ESRA transformer block serves as the feature extraction module in SUnet, and we stack two ESRA transformer blocks in each stage. The schematic diagram of the ESRA transformer block is illustrated in Fig. 1(b). The Mix-FFN module within the ESRA transformer block is depicted in Fig. 1(c), which differs from the traditional feed-forward network (FFN) via the use of depth-wise convolutions between the two linear layers. In the SUnet model, overlap patch embedding uses a convolutional layer with kernel size 7×7 and strides 3 for patch embedding. Overlap patch merging employs a convolutional layer with kernel size 3×3. The overlapping properties of the embedded patches help to mitigate the information loss caused by conventional patch embedding. Fig. 1(d) demonstrates the architecture of EFF, which mainly consists of three sub-modules: EAG, ECA and SA. We improve the original AG using grouped convolution with a group number of 32 and residual connections. The EAG module enhances the low-level semantic features transmitted through skip connections by the high-level semantic feature obtained through upsampling. The structure of EAG is presented in Fig. 1(e). After concatenation, ECA and SA are primarily used to highlight the important channels and spatial positions of task-relevant areas in the feature map to improve the ability of feature expression. It should be noted that the bottom ECA&SA module contains only a single input feature, so we only use ECA and SA for feature emphasis.

3.2. ESRA transformer block module

The transformer has gained significant popularity in computer vision tasks owing to its strong global modeling capabilities. However, when trained with limited data, transformer-based models often encounter challenges such as high computation complexity and susceptibility to overfitting. In order to address these challenges, we propose an approach called ESRA, which is depicted in Fig. 2.

The ESRA not only alleviates model overfitting but also reduces the overall parameters. Specifically, we utilize convolutional operations to compress the key and value in the MHSA, thereby decreasing the model parameters. The parameter count of the original MHSA can be expressed using Equation (1).

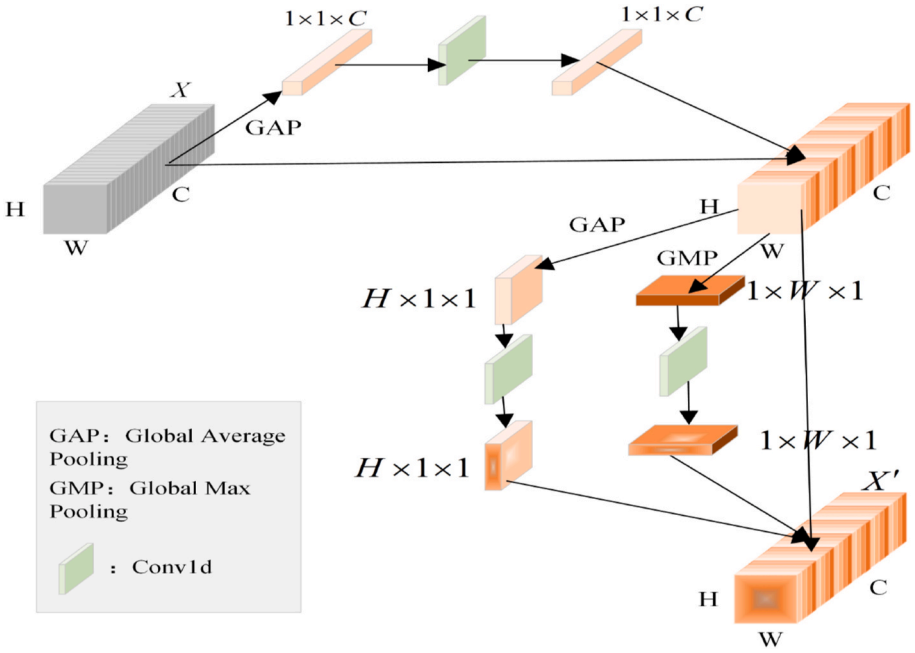

Fig. 3. ECA and SA structures in series. GAP is global average pooling, GMP is global maximum pooling, X is the input feature, $X^{'}$ is the output feature.

$$MHSA_{parameter} = head \times H \times W \times (4 \times dim_{head} + H \times W), \tag{1}$$

where H and W represent the height and width of the input feature map, *head* is the number of heads in the MHSA, and dim_{head} represents the channel dimension of each head. Here we default batch size to 1. The parameter calculation for ESRA follows the same principle and can be represented using Equation (2).

$$ESRA_{parameter} = head \times H \times W \times \frac{2 \times \left[dim_{head} \times \left(R_i^2 + 1\right) + H \times W\right]}{R_i^2}, \tag{2}$$

where, R_i represents the reduction ratio in the i-th stage. From Equation (1) and Equation (2), we can observe that when $R_i > 2$, the number of parameters of MHSA is greater than that of ESRA.

To mitigate the overfitting problem caused by the transformer, we employ dropkey [47] in ESRA to implicitly assign an adaptive operator to each attention head. This approach helps to constrain the attention distribution by penalizing regions with higher attention values, promoting smoother attention and encouraging the model to focus on other places relevant to the task, capturing robust global features. Therefore, ESRA can be represented as follows:

$$ESRA(Q, K, V) = Softmax\left(\frac{DropKey(Q * SR(K)^T)}{\sqrt{dim_{head}}} * SR(V)\right), \tag{3}$$

the $SR()$ operation can be expressed as follows :

$$SR(x) = Norm(Conv2d_{R_i}(x)), \tag{4}$$

in Equation (4), $Conv2d_{R_i}$ represents the feature compression achieved through a 2D convolution operation using a kernel size of R_i and a stride of R_i. And the dropkey operation can be expressed as follows:

$$DropKey(x) = x + bernoulli\left(ones\_like(x) * ratio_{dropkey}\right) * -e^{12}, \tag{5}$$

Where x represents the attention weights to be processed. The function *bernoulli*() is used to generate samples that follow a Bernoulli distribution, while *ones_like*() generates a matrix of ones with the same size as x.

3.3. Efficient feature fusion module

3.3.1. EAG module

The AG module was proposed by Oktay et al. in the Attention Unet [36]. However, we have found that the AG module not only suffers from high computational complexity but also requires strong data dependency. When AG is applied to high-resolution images, it increases the computational burden significantly. Moreover, there must be a rigorous data reliance between the two inputs of AG for it to capture important features accurately. When the correlation between the inputs is weak, AG fails to capture the crucial features. In our work, we extend the AG by replacing the conventional convolution with grouped convolution to conduct an intra-group feature fusion, and the calculation complexity of grouped convolution is obviously less than that of the conventional convolution method. At the same time, we modify the structure of AG by adding a ReLU layer after convolution of input features, and carrying out a residual connection for low-level semantic features passed by skip connections. Residual connections can mitigate the influence of high-level semantic features on low-level semantic features when the correlation between the two input features is weak, thereby avoiding performance degradation of the overall model. The internal structure of EAG is shown in Fig. 1(e). EAG can be expressed as follows:

$$EAG(g, x) = x \times \left(1 + Sigmoid\left(Conv_{1\times1}\left(ReLU\left(W_g + W_x\right)\right)\right)\right), \tag{6}$$

$$W_g = ReLU(BN(GroupConv_{32}(g))), \tag{7}$$

$$W_x = ReLU(BN(GroupConv_{32}(x))). \tag{8}$$

Where *Sigmoid* and *ReLU* are activation functions, *BN* is Batch Normalization operation, $GroupConv_{32}$ is grouped convolution with 32 groups, and $Conv_{1\times1}$ is conventional convolution with convolution kernel size of 1×1. In this model, g is the semantic feature obtained by up-sampling, and x is a low-level semantic feature passed by the skip connections.

3.3.2. EFF module based on multi-attention

The structure of the EFF module is shown in Fig. 1(d). In the EFF module, two semantic features from different levels are first enhanced by EAG to weaken the influence of unrelated regions. After the concatenation, the number of channels is double that of the original. As a large

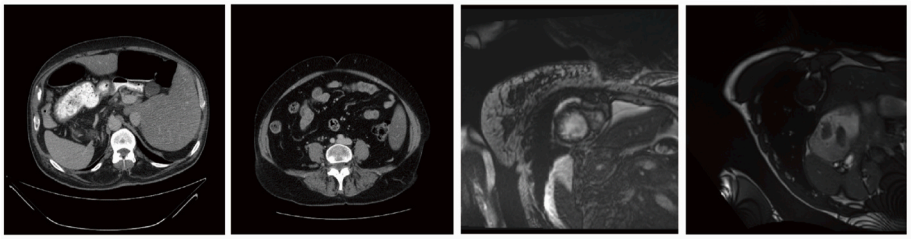

Fig. 4. Partial dataset images. (1) and (2) are images from Synapse, (3) and (4) are images from ACDC.

number of image information can be lost if operated directly, we use ECA and SA to emphasize related features from both dimensions. Notably, in our model, ECA and SA are connected in series. This combination of channel attention and spatial attention can better achieve multi-attention fusion. ECA and SA are used together and the structure diagram is presented in Fig. 3.

4. Experiments and results

In this section, we will discuss the experiments and results conducted to evaluate the performance of the proposed SUnet model. We compared its performance against benchmarked models through quantitative and qualitative experiments. Additionally, we performed structure ablation experiments to further analyze the model's architecture. These experiments were based on two public datasets. synapse multiple organ segmentation dataset (Synapse) [48] and automated cardiac diagnostic challenge dataset (ACDC) [49]. These datasets provided a reliable basis for assessing the effectiveness and generalizability of our model.

4.1. Datasets

Synapse Multiple Organ Segmentation Dataset (Synapse): In this experiment, we used 30 labeled abdominal CT scans and 3779 enhanced abdominal images from the MICCAI 2015 Multi-Atlas Abdominal Labeling Challenge. Each CT scan consists of 85–198 slices with a resolution of the 512 by 512 pixels. We performed image segmentation on eight different organs: aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen and stomach.

Automated Cardiac Diagnostic Challenge Dataset (ACDC): This dataset is widely used for cardiac magnetic resonance imaging (MRI) evaluation, providing a comprehensive and fully annotated collection of cardiac MRI scans. This dataset contained MRI scans of the hearts of 100 different patients, with each sample containing three organ tags, known as the left ventricle (LV), right ventricle (RV) and myocardium (Myo). In Fig. 4, partial dataset images were displayed.

4.2. Implementation details

All experiments in this paper are based on the Pytorch 1.8.0 framework [50]. We used a computer with Ubuntu 18.04 operating system, CPU I7-12700K, Nvidia RTX 3090, and 1 TB solid state drive to carry out the experiment. In all experiments of SUnet, we use the AdamW optimizer with the learning rate and weight decay are set to 1e-4.

For comparison purpose, we used the same hyperparameter setting in our model and all the benchmarked models. In the experiment of Synapse dataset, the data were divided into a training set consisting of 18 sample data, and a test set, consisting of 12 sample data. And we set the batch size to 24, the maximum number of epochs to 150, and the input image size and patch size to 224 × 224 and 16, respectively. Random flipping and rotation were applied to enhance the data.

In the experiment on the ACDC dataset, we use 70 scanned samples for training, 10 scanned samples for validation and 20 samples for testing. We set the batch size to 12, the epochs to 150, and patch size to 16. We used random flipping and rotation to enhance the data. Dice loss and Cross Entropy loss function were used, and the overall loss of the model could be expressed as:

$$LOSS = \lambda_1 \times DICE + \lambda_2 \times CE, \tag{9}$$

where, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$, *DICE* represents dice loss function, *CE* is cross entropy loss function.

4.3. Results on synapse and ACDC

To verify the model performance of SUnet, 10 representative 2D medical image segmentation models with good performance were selected for comparative test on the Synapse multi-organ segmentation dataset. The reason for selecting these models is their representativeness in the field. Unet is the pioneering U-shaped medical image segmentation network and forms the foundation of our model architecture. TransUnet is the first segmentation model that combines transformers with Unet. Swin Unet, on the other hand, is the first pure transformer-based U-shaped medical segmentation network. The remaining models

Table 1
Comparison of different methods in Synapse.

Methods	Parameters (M)	FLOPs (G)	DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
U-Net [22]	-	-	76.85	39.7	89.07	69.72	77.77	68.60	93.43	53.98	86.67	75.58
Att-Unet [36]	10.04	-	77.77	36.02	89.55	68.88	77.98	71.11	93.57	58.04	87.30	75.75
R50-ViT [22]	-	-	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUnet [22]	105.28	24.66	77.48	31.69	87.23.	63.13	81.87	77.02	94.08	55.86	85.08	75.62
SwinUnet [23]	27.17	5.90	79.13	21.55	85.47	66.53	83.28	79.61	94.29	56.58	90.66	76.60
PVTCASCADE [46]	34.13	5.84	81.06	20.23	83.01	70.59	82.23	80.37	94.08	64.43	90.10	83.6
MISSFormer [52]	42.46	7.21	81.96	18.2	86.99	68.65	85.21	82.00	94.41	65.67	91.92	80.81
TransCASCADE [46]	115.01	26.14	82.68	17.34	86.63	68.48	87.66	84.56	94.43	65.33	90.79	83.50
Parallel MERIT [51]	147.86	33.31	84.22	16.51	88.38	73.48	87.21	84.31	95.06	69.97	91.21	84.15
Cascaded MERIT [51] ★	147.86	33.31	84.90	13.22	87.71	74.40	87.79	84.85	95.26	71.81	92.01	85.38
SUnet (Ours)	**20.90**	**4.58**	**84.29**	**19.46**	**87.29**	**73.92**	**86.85**	**83.44**	**95.35**	**69.7**	**92.54**	**85.24**

Where ★ represents the current method of obtaining SOTA results. Our results are shown in bold. – indicates that the corresponding content is not found in the relevant paper. The index of each organ was average Dice Similarity coefficient (DSC).

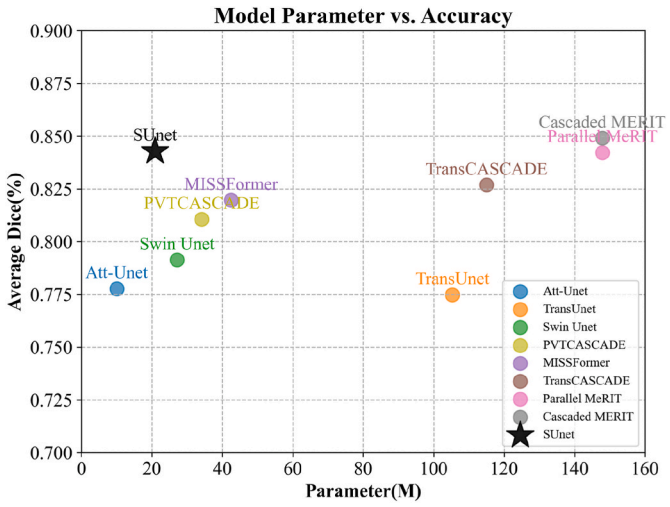

Fig. 5. Performance of 9 semantic segmentation models on Synapse. The X-axis represents the number of model parameters (unit: Million), and the Y-axis is average Dice, representing the performance of the model on the Synapse.

have all achieved top performance on the Synapse and ACDC datasets at different points in time.

Model evaluation data for U-Net, TransUnet, SwinUnet, TransCASCADE and other models were derived from published papers. The parameters and FLOPs are both obtained by the code they publish after the get_model_complexity_info function in the ptflops library. It should be noted that we failed to obtain the exact number of parameters and FLOPs of TransCASCADE due to the reason of TransCASCADE codes, so we used the number of parameters of PVTCASCADE minus the number of parameters of PVT model plus the number of parameters of TransUnet as the number of parameters of TransCASCADE. The same way calculated its FLOPs. Finally, statistical histograms and line graphs were used to represent the model performance visually.

4.3.1. *Result on synapse*

Table 1 presents the comparison test with ten deep learning models on the synapse multi-organ segmentation dataset. It is evident that the proposed SUnet model significantly outperforms the previous medical image segmentation models on the Synapse. Compared with the original Unet model, SUnet improved the Dice by 7.44%. Compared to the TransCASCADE model, it improved the Dice by 1.61% and compared to the PVTCASCADE model of the PVT model, it improved by 3.23%. At the same time, despite its superior performance, the SUnet model maintains a significantly smaller model scale compared to other transformer-based models, showcasing its efficiency and effectiveness.

At present, the state-of-the-art (SOTA) model in the 2D medical image segmentation field is the MERIT model proposed by Rahman et al. on March 29, 2023 [51]. The paper proposed two heterogeneous models, Cascaded MERIT and Parallel MERIT, where Cascaded MERIT achieved SOTA results on the Synapse multi-organ segmentation dataset. Although the performance of this paper is 0.61% lower than that of the Cascaded MERIT, it is important to highlight that the parameters and calculation of our model are 1/7 of its model scale. This demonstrates that our proposed SUnet model has better adaptability and more impressive performance for application scenarios with insufficient computing resources.

The scatter plots in Fig. 5 provide a visual comparison of 9 models, including Att-Unet, TransUnet, Swin-Unet, PVTCASCADE, MISSFormer, TransCASCADE and current SOTA models Parallel MeRIT and Cascaded MeRIT. The plots display the parameters and average Dice on the Synapse dataset. From the scatter plots, we can intuitively see that SUnet has fewer parameters than other models at the same level of approximate accuracy. With the same number of parameters, SUnet has a

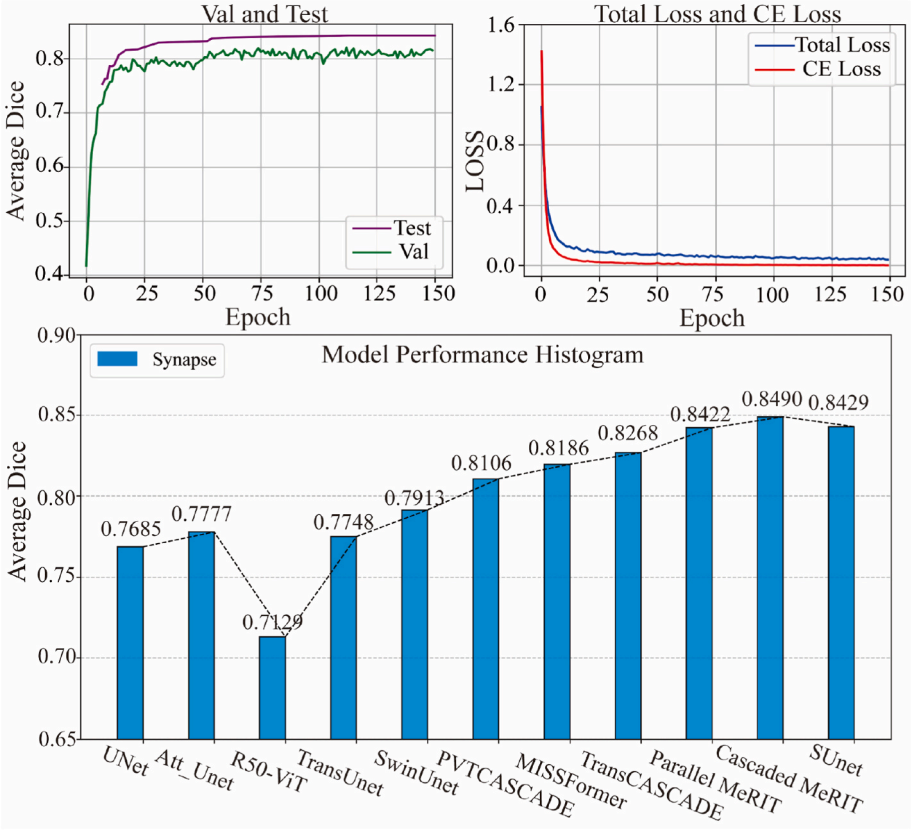

Fig. 6. Results of SUnet model on Synapse dataset.

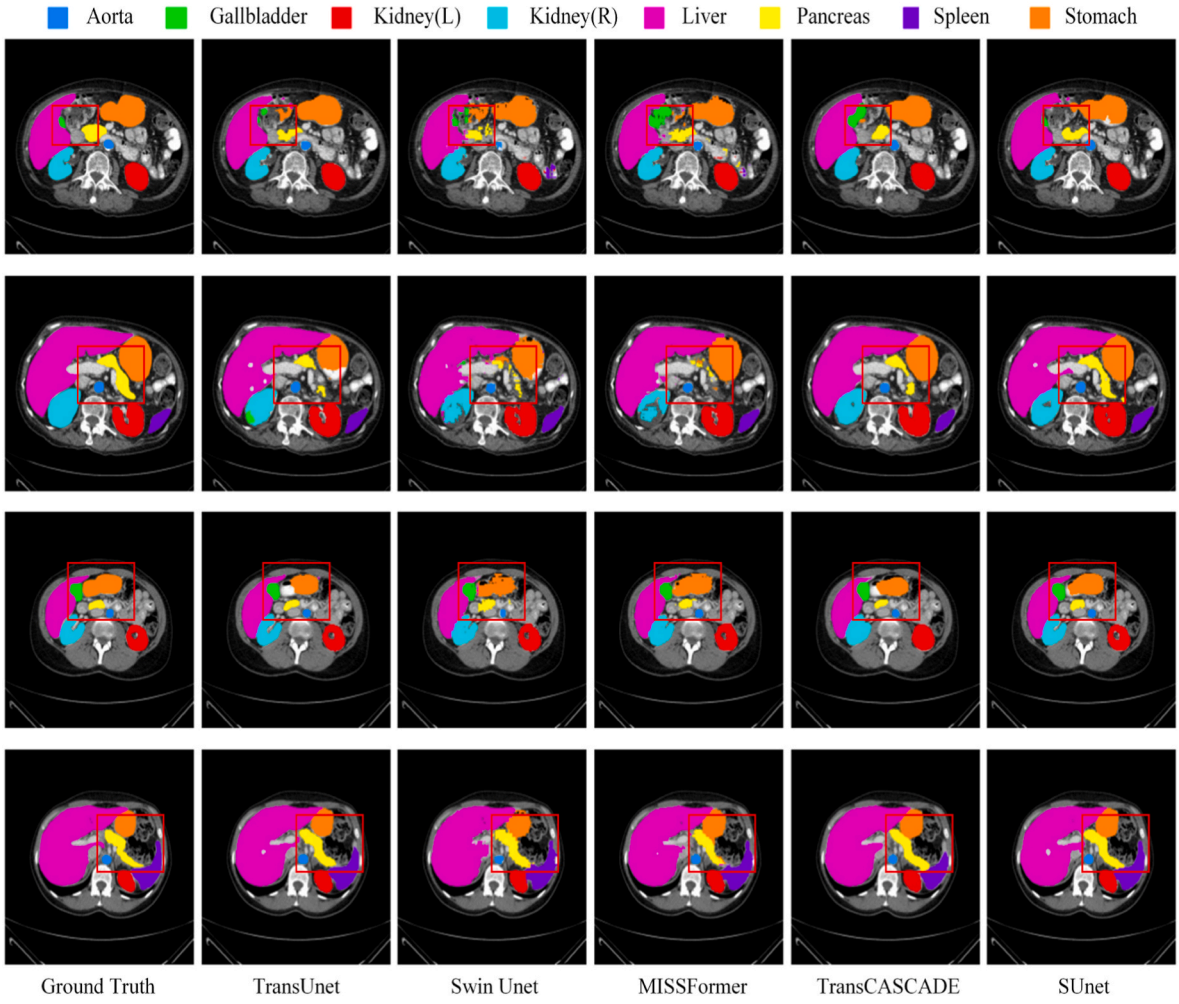

Fig. 7. Qualitative experiments of five models on the Synapse dataset. 8 different colors are used to represent the 8 labels to be divided and the segmentation results. We use red rectangular boxes to mark areas showing obvious differences between models. Each one lists the segmentation effect a model achieves on four random samples in the test set. Ground Truth is the result of expert segmentation.

Table 2
Comparison of different methods in ACDC.

Methods	Parameters (M)	FLOPs (G)	DSC↑	RV	Myo	LV
U-Net [22]	-	-	87.55	87.10	80.63	94.92
Att-Unet [36]	10.04	-	86.75	87.58	79.20	93.47
R50-ViT [22]	-	-	87.57	86.07	81.88	94.75
TransUnet [22]	105.28	24.66	89.71	88.86	84.53	95.73
SwinUnet [23]	27.17	5.90	90.00	88.55	85.62	95.83
PVTCASCADE [46]	34.13	5.84	91.46	88.90	89.97	95.50
MISSFormer [52]	42.46	7.21	90.86	89.55	88.04	94.99
TransCASCADE [46]	115.01	26.14	91.63	89.14	90.25	95.50
Parallel MERIT [51]	147.86	33.31	92.32	90.87	90.00	96.08
Cascaded MERIT [51]	147.86	33.31	91.85	90.23	89.53	95.80
SUnet (Ours)	**20.90**	**4.58**	**92.25**	**90.69**	**89.95**	**96.09**

significant advantage in terms of model accuracy. This proves that our model is well-suited for the environments with limited memory and computing resources. The convergence curve of the loss and the Dice growth curve of SUnet on the Synapse dataset are shown in Fig. 6.

Furthermore, in Fig. 7, we provide qualitative analysis results of the Synapse dataset for representative models, including TransUnet, SwinUnet, MISSFormer, TransCASCADE, and SUnet. We randomly selected four samples in the test set for comparison experiment. The qualitative experiment shows that the SUnet model outperforms the other four models regarding segmentation performance for all four samples. This qualitative analysis further supports the superior performance of our model in terms of segmentation accuracy.

4.3.2. *Result on ACDC*

Meanwhile, we conducted comparative tests on the ACDC dataset, as illustrated in Table 2. In the ACDC dataset comparison experiment, we compared the same ten medical image segmentation models as in the Synapse comparison experiment. Compared to TransCASCADE models, our model demonstrated an improvement of 0.62%. Although the performance improvement of SUnet on the ACDC data set for the model is not considerable, it is still a significant breakthrough to achieve such excellent results without a large increase in the parameters. The convergence curve of the loss and the Dice growth curve of SUnet on the ACDC dataset are shown in Fig. 8. The visualized result is shown in Fig. 9.

Our model shows a 0.4% performance advantage over cascading MERIT on the ACDC dataset. However, compared to the parallel MERIT, SUnet was 0.07% behind in average Dice. It is important to note that the SUnet model is a single structural model. In contrast, the cascade MERIT and parallel MERIT models proposed by Rahman are two heterogeneous

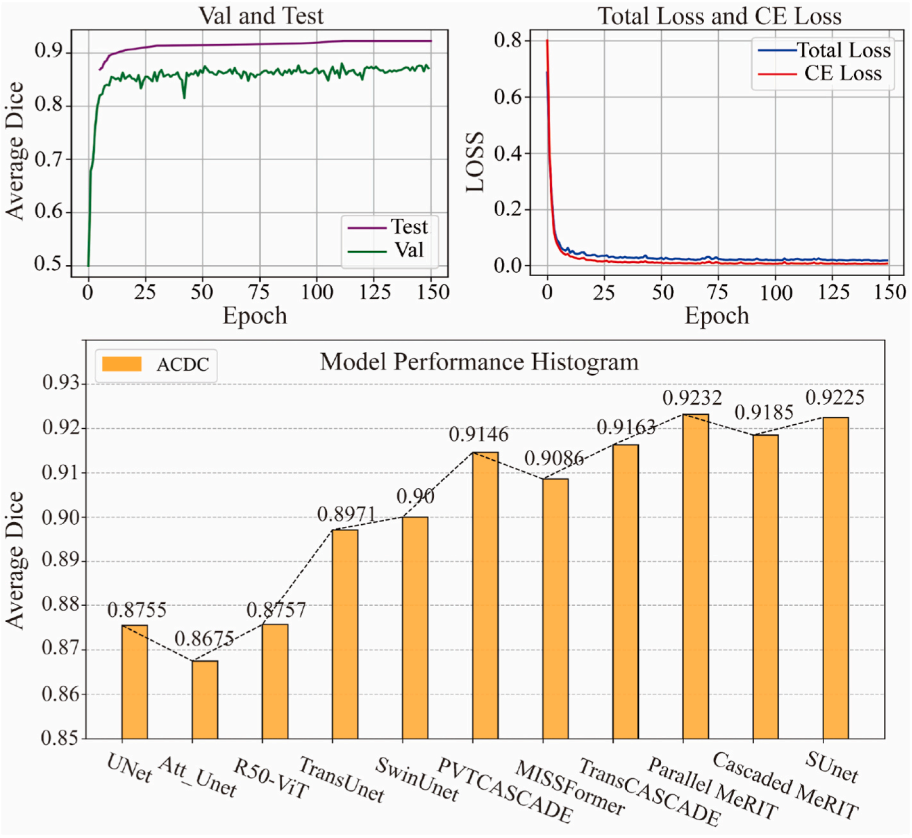

Fig. 8. Results of SUnet model on ACDC dataset.

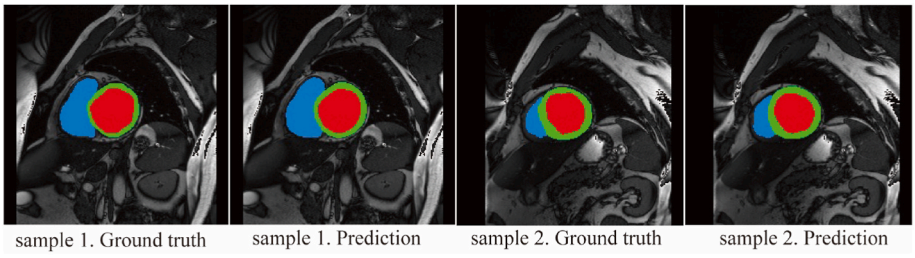

Fig. 9. Visualization results of SUnet model on ACDC dataset.

Table 3
Ablation experiment on Synapse.

Methods	ESRA	EFF			DSC↑	HD↓	Aorta	Gallbladder	Kidney(L)	Kidney(R)	Liver	Pancreas	Spleen	Stomach
		EAG	ECA	SA										
SUnet-0	✓	x	x	x	82.86	20.21	87.69	73.42	86.69	81.59	95.57	65.21	91.27	81.47
SUnet-1	✓	✓	x	x	83.65	17.02	87.48	72.60	87.05	83.62	95.75	67.90	90.70	84.11
SUnet-2	✓	✓	✓	x	81.72	25.42	87.83	69.78	82.49	79.34	94.97	66.06	89.86	83.47
SUnet-3	✓	✓	x	✓	81.79	27.73	87.60	74.31	80.81	78.71	95.42	69.61	87.97	78.89
SUnet-4	✓	x	✓	✓	83.29	20.31	88.10	72.34	84.62	82.46	95.40	70.00	90.65	82.73
SUnet	✓	✓	✓	✓	84.29	19.46	87.29	73.92	86.85	83.44	95.35	69.7	92.54	85.24

medical image segmentation models. Furthermore, the SUnet model has significantly fewer parameters and FLOPs than the Cascade MERIT and Parallel MERIT models. From this point of view, our method can demonstrate better performance in scenarios with limited computational resources.

4.4. Ablation experiment

The paper conducted ablation experiments to validate the effectiveness of the proposed ESRA and EFF modules. A series of structural ablations were performed on the three sub-modules in EFF to evaluate their individual contributions. Firstly, we conducted experiments by using only ESRA to construct SUnet-0 as the baseline model. SUnet-1 is built upon SUnet-0 by incorporating the EAG module. The remaining

structural ablations are shown in Table 3, and all the ablation experiments were conducted on the abdominal multi-label dataset, while keeping all other aspects consistent except for the model structure.

The results of the ablation experiments are shown in Table 3. It is evident that the SUnet-0, which is solely based on ESRA, achieves an average Dice of 82.86% on the Synapse, surpassing the performance of most 2D medical image segmentation models. And the performance of the SUnet-1 model, which incorporates the EAG module on top of SUnet-0, is further improved. We also observed an interesting phenomenon that introducing only ECA or SA on the SUnet-1 decreased model performance. SUnet-2 and SUnet-3 achieved average Dice of 81.72% and 81.79%, respectively. But, when SUnet-4 utilized ESRA along with the concatenated ECA and SA, the model's performance improved by 0.43% compared to SUnet-0. This suggests that if the model excessively emphasizes a distinct dimension, it may become trapped in a single-dimensional local optimum. Consequently, we believe it is necessary to consider ECA for channel attention and SA for task-specific region attention as a unified approach. In conclusion, the results from the ablation experiments visually demonstrated that the proposed methods in this study are effective in significantly enhancing the model's performance.

5. Conclusion

In this paper, we introduce SUnet, a novel 2D medical image segmentation model based on the ESRA. We propose an innovative EFF module that effectively fuses skip connections and decoder features using multiple attention mechanisms, including EAG, ECA, and SA. The EAG module, based on grouped convolution, enables efficient intra-group feature fusion. Compared to other 2D medical image segmentation models, such as TransUnet and Swin Unet, our proposed SUnet model achieves higher accuracy with fewer parameters. It achieves an average Dice of 84.29% on the Synapse dataset and 92.25% on the ACDC dataset. SUnet demonstrates superior adaptability and parameter efficiency over current state-of-the-art 2D medical image segmentation models, making it more suitable for various tasks in 2D medical image segmentation, particularly in scenarios with limited computational resources.

Notwithstanding the use of multiple attention mechanisms to achieve feature fusion, SUnet has not effectively solved the problem of fusing local and global features at a fundamental level. In future research, we will develop a new efficient semantic segmentation model that integrates global and local image features to extract more effective image features for improved medical image segmentation. For example, we aim to form a new hybrid feature extraction unit that pays equal attention to global and local features by using CNN and transformer. The challenge lies in achieving this more efficient hybrid model while ensuring that it does not significantly increase the parameters or even become more lightweight. By addressing these challenges, we aspire to enhance the performance of medical image segmentation models and contribute to computer-aided diagnosis and treatment.

Funding

This work is supported by the Ministry of Science and Technology of the People's Republic of China under Grant No. STI2030-Major Projects 2021ZD0201900, the National Natural Science Foundation of China under Grant No. 12090052, Guangxi Key R&D Project under Grant No. AB21076021, AA22068057, and supported by Open Research Fund of Guangxi Key Lab of Human-machine Interaction and Intelligent Decision (GXHIID2207) and Center for Applied Mathematics of Guangxi (Nanning Normal University).

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] H. Gao, M. Lyu, X. Zhao, et al., Contour-aware network with class-wise convolutions for 3D abdominal multi-organ segmentation, Med. Image Anal. 87 (2023), 102838.
[2] H. Hu, Z. Feng, H. Lin, et al., Gene function and cell surface protein association analysis based on single-cell multiomics data, Comput. Biol. Med. 157 (2023), 106733.
[3] H. Hu, Z. Feng, H. Lin, et al., Modeling and analyzing single-cell multimodal data with deep parametric inference, Briefings Bioinf. 24 (2023), bbad005.
[4] W. Wang, L. Zhang, J. Sun, et al., Predicting the potential human lncRNA–miRNA interactions based on graph convolution network with conditional random field, Briefings Bioinf. 23 (2022) bbac463.
[5] H. Zhang, Y. Wang, Z. Pan, et al., ncRNAInter: a novel strategy based on graph neural network to discover interactions between lncRNA and miRNA, Briefings Bioinf. 23 (2022) bbac411.
[6] S. Zhang, K. Amahong, C. Zhang, et al., RNA–RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection, Briefings Bioinf. 23 (2022) bbab397.
[7] K. Amahong, W. Zhang, Y. Zhou, et al., CovInter: interaction data between coronavirus RNAs and host proteins, Nucleic Acids Res. 51 (2023) D546–D556.
[8] J. Fu, Q. Yang, Y. Luo, et al., Label-free proteome quantification and evaluation, Briefings Bioinf. 24 (2023) bbac477.
[9] J. Zhao, J. Sun, S.C. Shuai, et al., Predicting potential interactions between lncRNAs and proteins via combined graph auto-encoder methods, Briefings Bioinf. 24 (2023) bbac527.
[10] F. Xu, D. Miao, W. Li, et al., Specificity and competition of mRNAs dominate droplet pattern in protein phase separation, Phys. Rev. Res. 5 (2023), 023159.
[11] Q. Yang, Y. Gong, F. Zhu, Critical assessment of the biomarker discovery and classification methods for multiclass metabolomics, Anal. Chem. 95 (2023) 5542–5552.
[12] X. Li, P. Zhang, Z. Yin, et al., Caspase-1 and Gasdermin D afford the optimal targets with distinct switching strategies in NLRP1b Inflammasome-induced cell death, Research 2022 (2022), 9838341.
[13] X. Li, C.-Q. Zhong, R. Wu, et al., RIP1-dependent linear and nonlinear recruitments of caspase-8 and RIP3 respectively to necrosome specify distinct cell death outcomes, Protein Cell 12 (2021) 858–876.
[14] J. Fu, Y. Zhang, J. Liu, et al., Pharmacometabonomics: data processing and statistical analysis, Briefings Bioinf. 22 (2021) bbab138.
[15] 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.
[16] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, et al., Unet++: redesigning skip connections to exploit multiscale features in image segmentation, IEEE Trans. Med. Imag. 39 (2019) 1856–1867.
[17] H. Huang, L. Lin, R. Tong, et al., Unet 3+: A Full-Scale Connected Unet for Medical Image Segmentation, ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1055–1059.
[18] D. Jha, P.H. Smedsrud, M.A. Riegler, et al., Kvasir-seg: A Segmented Polyp Dataset, MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26, Springer, 2020, pp. 451–462.
[19] D. Jha, P.H. Smedsrud, M.A. Riegler, et al., Resunet++: an advanced architecture for medical image segmentation, in: 2019 IEEE International Symposium on Multimedia (ISM), IEEE, 2019, pp. 225–2255.
[20] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, Proc. Adv. Neural Inf. Process. Syst. 30 (2017).
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020 arXiv preprint arXiv: 2010.11929.
[22] J. Chen, Y. Lu, Q. Yu, et al., Transunet: Transformers Make Strong Encoders for Medical Image Segmentation, 2021 arXiv preprint arXiv:2102.04306.
[23] H. Cao, Y. Wang, J. Chen, et al., Swin-unet: Unet-like Pure Transformer for Medical Image Segmentation, European Conference on Computer Vision, Springer, 2022, pp. 205–218.
[24] N.T. Duc, N.T. Oanh, N.T. Thuy, et al., Colonformer: an efficient transformer based method for colon polyp segmentation, IEEE Access 10 (2022) 80575–80586.
[25] K. Fitzgerald, B. Matuszewski, FCB-SwinV2 Transformer for Polyp Segmentation, 2023 arXiv preprint arXiv:2302.01027.
[26] E. Sanderson, B.J. Matuszewski, FCN-transformer feature fusion for polyp segmentation, in: Annual Conference on Medical Image Understanding and Analysis, Springer, 2022, pp. 892–907.
[27] Z. Li, Y. Li, Q. Li, et al., Lvit: language meets vision transformer in medical image segmentation, IEEE Trans. Med. Imag. (2023) 1. -1.
[28] L. Zhou, Spatially Exclusive Pasting: A General Data Augmentation for the Polyp Segmentation, 2022 arXiv preprint arXiv:2211.08284.
[29] J. Wang, Q. Huang, F. Tang, et al., Stepwise Feature Fusion: Local Guides Global, International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 110–120.
[30] Y. Zhang, H. Liu, Q. Hu, Transfuse: fusing transformers and cnns for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2021, pp. 14–24.

[31] J. Wang, F. Chen, Y. Ma, et al., XBound-Former: toward cross-scale boundary modeling in Transformers, IEEE Trans. Med. Imag. 42 (2023) 1735–1745.
[32] W. Wang, E. Xie, X. Li, et al., Pyramid vision transformer: a versatile backbone for dense prediction without convolutions, Proceedings of the IEEE/CVF international conference on computer vision (2021) 568–578.
[33] P.K. Buttar, M.K. Sachan, Semantic segmentation of clouds in satellite images based on U-Net++ architecture and attention mechanism, Expert Syst. Appl. 209 (2022), 118380.
[34] M. Wieland, S. Martinis, R. Kiefl, et al., Semantic segmentation of water bodies in very high-resolution satellite and aerial images, Remote Sens. Environ. 287 (2023), 113452.
[35] J. Jiang, J. Zhu, M. Bilal, et al., Masked swin transformer unet for industrial anomaly detection, IEEE Trans. Ind. Inf. 19 (2022) 2200–2209.
[36] O. Oktay, J. Schlemper, L.L. Folgoc, et al., Attention U-Net: Learning where to Look for the Pancreas, 2018 arXiv preprint arXiv:1804.03999.
[37] Z. Han, M. Jian, G.-G. Wang, ConvUNeXt: an efficient convolution neural network for medical image segmentation, Knowl. Base Syst. 253 (2022), 109512.
[38] X. Li, H. Chen, X. Qi, et al., H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, IEEE Trans. Med. Imag. 37 (2018) 2663–2674.
[39] S. Roy, G. Koehler, C. Ulrich, et al., MedNeXt: Transformer-Driven Scaling of ConvNets for Medical Image Segmentation, 2023 arXiv preprint arXiv:2303.09975.
[40] F. Isensee, J. Petersen, A. Klein, et al., nnu-net: Self-adapting framework for u-net-based medical image segmentation (2018) arXiv preprint arXiv:1809.10486.
[41] B. Li, S. Liu, F. Wu, et al., RT-Unet: an advanced network based on residual network and transformer for medical image segmentation, Int. J. Intell. Syst. 37 (2022) 8565–8582.
[42] H. Wang, Y. Fan, Z. Wang, et al., Parameter-free Spatial Attention Network for Person Re-identification, 2018 arXiv preprint arXiv:1811.12150.
[43] 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.
[44] Q. Wang, B. Wu, P. Zhu, et al., ECA-Net: efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.
[45] S. Woo, J. Park, J.-Y. Lee, et al., Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[46] M.M. Rahman, R. Marculescu, Medical image segmentation via cascaded attention decoding, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2023) 6222–6231.
[47] B. Li, Y. Hu, X. Nie, et al., DropKey, 2022 arXiv preprint arXiv:2208.02646.
[48] B. Landman, Z. Xu, J. Igelsias, et al., Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge, in: Proc. MICCAI Multi-Atlas Labeling beyond Cranial Vault—Workshop Challenge, 2015, p. 12.
[49] O. Bernard, A. Lalande, C. Zotti, et al., Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imag. 37 (2018) 2514–2525.
[50] A. Paszke, S. Gross, F. Massa, et al., Pytorch: an imperative style, high-performance deep learning library, Proc. Adv. Neural Inf. Process. Syst. 32 (2019).
[51] M.M. Rahman, R. Marculescu, Multi-scale Hierarchical Vision Transformer with Cascaded Attention Decoding for Medical Image Segmentation, 2023 arXiv preprint arXiv:2303.16892.
[52] X. Huang, Z. Deng, D. Li, et al., Missformer: an Effective Medical Image Segmentation Transformer, 2021 arXiv preprint arXiv:2109.07162.