Biomedical Signal Processing and Control 102 (2025) 107203

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

SEAformer: Selective Edge Aggregation transformer for 2D medical image segmentation

Jingwen Li [a], Jilong Chen [b,c], Lei jiang [b,c], Ruoyu Li [b], Peilun Han [d,e], Junlong Cheng [b,c,*]

[a] School of Computer Science and Technology, Xinjiang University, Urumqi, Xinjiang, 830000, China
[b] School of Computer Science, Sichuan University, Chengdu, Sichuan, 610065, China
[c] Vision Computing Lab, Sichuan University, Chengdu, Sichuan, 610065, China
[d] West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China
[e] Department of Radiology, West China Hospital, Sichuan University, Chengdu, Sichuan, 610041, China

ARTICLE INFO

Keywords:
Medical image segmentation
Transformer
Densely connected
Selective edge aggregation
Multi-level optimization strategy

ABSTRACT

Automatic medical image segmentation has a wide range of applications and high research values in medical research and practice, which can assist medical workers in clinical lesion assessment and diagnosis analysis of disease. However, it is still challenging due to the large-scale variations, blurred structural boundaries, and irregular shapes of segmentation targets in medical images. To tackle these challenges, we propose a selective edge aggregation Transformer (SEAformer) with an encoder–decoder architecture for 2D medical image segmentation. Specifically, we first combine densely connected CNNs and Transformers (with Dense MLP) in a parallel manner to form a dual encoder that efficiently captures shallow texture information and global contextual information in medical images in a deeper, multi-scale way. Then, we propose a plug-and-play selective edge aggregation (SEA) module that removes the noisy background unsupervisedly, selects and retains useful edge features, making the network more focused on the information related to the target boundary. Finally, we design a loss function that combines the target content and edges and use a multi-level optimization (MLO) strategy to refine the blur structure. This optimization helps the network to learn better feature representations and produce more accurate segmentation results. In addition, due to our densely connected approach to building the entire network, SEAformer has only 16 MB parameters and 32 GFlops. Extensive experimental results show that SEAformer performs well compared with state-of-the-art methods in four different challenging medical segmentation tasks, including skin lesion segmentation, thyroid nodules segmentation, GLAnd segmentation, and COVID-19 infection segmentation.

1. Introduction

Medical image segmentation is a widely-studied and challenging topic, which aims to help clinicians focus more on pathological regions and extract detailed information from medical images for more accurate diagnosis and analysis. Current common medical image segmentation tasks include skin lesion segmentation, gland segmentation, thyroid nodule segmentation, etc [1]. However, the large-scale variations of the targets to be segmented in medical images (Fig. 1(a) shows different scale lesion/pathology images), blurred target structural boundaries (the green curve in Fig. 1(b) outlines the visual difficulty in identifying target boundaries), numerous modalities (Fig. 1(a) and (b) contains four different imaging modalities and medical images with different lesion shapes) and the lack of high-quality labeled images for practice training make it very difficult to obtain accurate segmentation results.

With the rapid development of deep learning techniques, many end-to-end automatic segmentation methods have been proposed and applied to medical image analysis [2,3]. U-Net [3] is one of the most widely used medical image segmentation models, which uses an encoder to learn advanced semantic representations, a decoder to recover lost spatial information, and applies a skip connection to fuse different scale features of the encoder and decoder to produce a more accurate segmentation mask. Subsequently, researchers have proposed many variants to improve U-Net, for example, (1) enhance the feature representation of the model by building deeper network structures or extracting rich multi-scale information [4,5]; (2) enhance the utilization of features at different levels by using nested sub-networks or multiplexing features [6–8]; (3) produce more discriminative feature representations through attention mechanisms [9,10]; (4) Encoding the long-range dependencies of an image via a Transformer or a hybrid

* Corresponding author.
E-mail address: cjl951015@163.com (J. Cheng).

https://doi.org/10.1016/j.bspc.2024.107203
Received 8 July 2024; Received in revised form 29 October 2024; Accepted 15 November 2024
Available online 4 December 2024

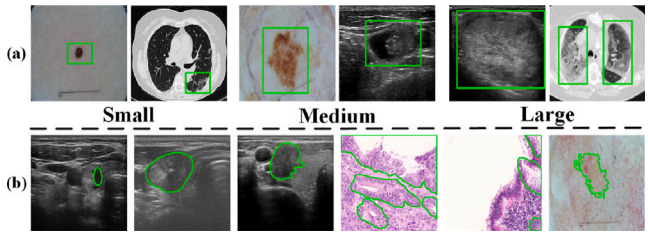

Fig. 1. (a) Lesion/pathology images at different target scales. (b) The case of blurred structural boundaries.

model [11–14]. However, these deep learning-based medical image segmentation methods do not explicitly take accurate boundary prediction into consideration which can produce higher-quality segmentation masks. To solve the problem of ill-defined structural boundaries, several methods have been reported [15–19] that recover boundary details by learning inter-pixel dependencies. However, they either require manual tuning of parameters during post-processing or elaborate learnable modules to perform this labor-intensive task.

Most existing convolutional neural networks (CNNs) methods are unable to establish long-term dependencies and global contextual connections due to the limitations of receptive fields in convolutional operations. Repeated stride and pooling operations inevitably lose the images' resolution making the dense prediction task challenging. This problem was greatly alleviated by the emergence of the Transformer, which was first used for natural language processing tasks and could encode long-range dependencies. ViT [20] is the first purely visual Transformer framework for image classification. Notably, PoolFormer [21] demonstrates that the success of Transformers and their variants largely stems from their overall architectural design, and substantial results can be achieved by replacing the token mixer with simple pooling operations. However, the good performance of Transformers often requires pretraining on large datasets, which limits the application of pure Transformers to medical image datasets with smaller amounts of data. Therefore, the fusion of CNNs and Transformers has become an important research direction. Transformers are better at handling non-local interactions, which is helpful for capturing the associations between long-distance pixels, a particularly important aspect for the complex structures and scattered information in medical images. Meanwhile, CNNs excel at extracting local features and texture information, and the fusion of the two can fully leverage their respective advantages. For example, TransU-Net [11] utilizes Transformers to optimize the feature extraction process of CNNs, while FATNet [22] and X-Net [23], among others, integrate CNN and Transformer branches in a parallel manner to achieve more accurate medical image segmentation. Our SEAformer adopts a similar fusion method with a parallel dual-encoder structure, which facilitates the simultaneous capture of local and global contextual information in medical images.

To address the aforementioned issues, we propose SEAformer for two-dimensional medical image segmentation. Firstly, it fully integrates the advantages of CNNs and Transformers by running in parallel encoder branches based on densely connected CNNs and Transformers. This not only encodes input information simultaneously but also allows the fusion and interaction of features at the same resolution. Constructing the encoder in this way offers the following advantages: (1) It can effectively capture global contextual features and shallow spatial features in medical images; (2) The densely connected approach provides natural multi-scale capability without the need to build additional functional modules with multi-scale capabilities. Subsequently, we introduce the Selective Edge Aggregation (SEA) module, which receives feature information from both the Transformer and CNN branches. The design intention of the SEA module is to leverage the spatial information captured by CNNs to achieve fusion and complementarity with the global features of Transformers. We integrate the SEA module into the Transformer architecture and replace the standard multi-head self-attention module, enabling adaptive removal of noisy backgrounds and selective retention of edge-related features, thereby effectively improving edge detection and segmentation accuracy. Meanwhile, the SEA module requires no additional learned parameters, thus reducing the computational burden of the secondary complexity of the original Transformer architecture. Finally, we design a loss function combining target edges and content based on the principle of SEA and adopt a multi-level optimization strategy to optimize the encoder and decoder. This strategy enables the model to comprehensively consider feature information at different scales, better understand the context and structure of the input data, and thus improves the model's ability to recognize and locate targets. Thanks to a codec architecture that is densely connected, SEAformer saves 45% of the number of parameters and 23% of the GFlops compared with the baseline U-Net [3]. We conducted extensive experiments on five medical segmentation datasets and performed a comprehensive analysis from both quantitative and qualitative perspectives. In parallel, we performed a series of ablation studies to assess the effectiveness of each component of the proposed method. The experimental results showed that SEAformer performed better compared with most state-of-the-art methods. Our main contributions can be summarized as follows:

- We propose SEAformer, a framework that effectively integrates the strengths of CNNs and Transformers. Unlike conventional sequential or parallel CNN-Transformer architectures, SEAformer simultaneously captures shallow spatial features and global contextual information, enabling thorough feature fusion and interaction across stages. The densely connected architecture endows the encoder–decoder with natural multi-scale learning capability while reducing parameter count. This design allows SEAformer to perform robustly across different medical imaging modalities, addressing challenges such as scale variation and blurred structural boundaries, even on small datasets.

- We design the SEA module to selectively aggregate edge-related features while effectively filtering background noise, significantly enhancing the network's ability to recognize precise edge features. Unlike prior methods relying on complex post-processing, the SEA module requires no additional learnable parameters, reducing computational complexity. We embed this module into the Transformer architecture, linking features from both CNN and Transformer branches.

- We introduce a loss function that combines edge and content features, specifically targeting boundary detail learning. The multi-level optimization strategy effectively fine-tunes the encoder and decoder, encouraging the network to capture more accurate and comprehensive boundary information. This targeted optimization produces improved feature representations and overall segmentation quality, demonstrating significant advantages over traditional loss functions.

2. Related work

2.1. Medical image segmentation based on CNNs

Fully Convolutional Network (FCN) [2] is the first end-to-end method of semantic segmentation that can accept an input image of arbitrary size and recover the feature map of the last convolutional layer to the size of the input image by interpolating up-sampling to produce a prediction for each pixel. Researchers have improved FCN for medical image segmentation [24,25] according to different task requirements. For example, Patrick et al. [24] combined cascaded fully convolutional networks (CFCN) and dense 3D conditional random fields

(CRF) to achieve automatic segmentation of liver and lesion regions in CT abdominal images. Chen et al. [25] used 3D V-Net for end-to-end lung segmentation while refining the V-Net output based on a priori shape knowledge using a deformation module. U-Net [3] is another encoder–decoder framework commonly used in medical images, which adds skip connections to FCN to fuse shallow features of the encoder and deep features of the decoder. Inspired by U-Net, researchers have proposed many variant methods based on CNNs [4–10,26,27]. Simon et al. [5] proposed a hybrid densely connected U-Net for automatic segmentation in the liver and tumor and cheng et al. [4] extended the densely connected architecture to the dual encoder structure. The literature [6–8]improved the segmentation performance by constructing recursive structures in U-Net. Zhang et al. [26] proposed MSDANet, which utilizes a parallel dilated pooling module to reduce the loss of subtle features during downsampling. Additionally, a multi-scale channel attention mechanism, a multilayer perceptron squeeze-and-excitation module, and a large kernel convolution module are employed in both the encoding and decoding stages to enhance the extraction of effective features. Gu et al. [27] introduced the atrous spatial pyramid pooling module into the U-Net structure to capture multi-scale information in medical images. Zhu et al. [28] introduces a multimodal spatial information enhancement and boundary shape correction method, comprising a modality information extraction module, a spatial information enhancement module, and a boundary shape correction module, to tackle the problem of end-to-end 3D brain tumor segmentation. Other researchers have enhanced the weight of important features and further improved the feature representation by integrating spatial/channel attention modules in U-Net [9,10]. These methods have achieved reasonable segmentation results in medical image segmentation, but they are still deficient in establishing long-range dependencies due to the limitations of receptive fields in convolutional operations.

2.2. Medical image segmentation based on vision transformer

The successful application of ViT [20] on ImageNet demonstrates the potential of pure Transformer for vision tasks. To perceive the global information of medical images, Chen et al. [11] integrated Transfomer into CNNs as an encoder and recovered local spatial information by up-sampling of U-Net to complete fine medical image segmentation. Tang et al. [29] designed a trident multi-layer fusion module for the Transformer to dynamically integrate contextual information from higher-level (global) features. Additionally, they developed united attention modules to focus on the learning of significant features. Jeya et al. [12] proposed a gated axial-attention model, which operated on the whole image and patches separately to learn global and local features. Zhu et al. [13] proposed SDV-TUNet, which utilizes a sparse dynamic encoder–decoder to extract global spatial semantic features for brain tumor segmentation. H2Former [30] and SegNetr [31] propose hierarchical hybrid vision Transformers designed for medical image segmentation, which can efficiently utilize data in limited medical data environments. Wu et al. [22] proposed a two-encoder segmentation network combining CNNs and Transformer and recovered high-resolution features from both branches using an efficient decoder and a feature adaptation module. Similarly, [32,33] combined the respective strengths of CNNs and Transformers to efficiently capture global dependencies and low-level spatial details in a more shallow manner and used the fusion module to efficiently fuse multi-level features from both branches. Xu et al. [14] proposed the backbone hybrid network and the deep micro-texture extraction Module based on the Swin-Transformer architecture. These modules extract global and local features from multimodal data, enabling them to capture and analyze deep texture features in multimodal images. Li et al. [23] proposed a dual encoding-decoding structure for the X-shaped network (X-Net). In the encoding stage, local and global features are simultaneously extracted by two types of encoders: a CNNs downsampling encoder and a Transformer encoder, and then fused through skip connections. We can find that although these approaches have successfully applied the vision Transformer to medical image segmentation, CNNs also play an important role in it. We believe that simply running the Transformer and CNNs in series or parallel does not fully exploit the advantages of both. Similar to FATNet [22] and TransFuse [32], SEAformer has a parallel dual encoder structure that captures both local and global contextual information in medical images. Unlike them, we have information interaction at each stage of Transformer and CNNs, which effectively ensures the information flow between the two branches and learns the different features while enhancing the weights of similar features.

2.3. Edge-awareness-based networks

Modeling with edge information of medical images helps to produce finer segmentation results. Several methods have been reported recently [15–19]. For example, Chen et al. [15] combined CNNs and probabilistic graph models to improve the localization of object boundaries and reprocessed the initial coarse segmentation results using fully connected conditional random fields. Zhu et al. [16] proposed a brain tumor segmentation method based on the fusion of deep semantics and edge information in multimodal MRI, and specially designed an edge space attention block for feature enhancement. Lee et al. [17] proposed a structural boundary-preserving segmentation framework to predict the structural boundary of a target object by estimating the key points on the structural boundary of the target object. In addition, they used a shape boundary-aware evaluator to provide feedback to the segmentation network based on the structural boundary key points. Similarly, BAT [18] used boundary prior knowledge to capture more local details in medical images, effectively dealing with fuzzy boundaries. EANet [19] constructed a dynamic scale perception context module to efficiently extract multi-scale contextual information and combined it with an edge attention-preserving module to remove noise and help edge streams focus on processing boundary-related information only. However, the applicability of these methods is somewhat limited due to their high computational cost, complicated tuning parameters, and fixed collocation.

In contrast, we design a selective edge aggregation module and integrate it into Transformer. This module selectively aggregates edge information using spatial feature maps extracted by CNNs, allowing the network to focus significantly on the boundaries of the target region. It is important to note that the SEA module allows adaptive selective aggregation by inputting feature maps without manually adjusting parameters. In addition, the module is a completely non-parametric module that can be integrated into any end-to-end segmentation network.

3. SEAformer

An overview of our proposed SEAformer framework for 2D medical image segmentation is shown in Fig. 2. The framework is a U-Net variant based on an encoder–decoder architecture and consists of three parts: a Transformer-based encoder, a densely connected CNN-based encoder, and a decoder. We capture and retain important global contextual information through the Transformer branch and extract rich local and spatial texture information using the CNN branch. Note that the dense connection has natural multi-scale feature extraction capability, and the encoder performs information interaction at each stage of both branches to fully exploit both local and global information. The densely connected CNN decoder fuses low-to-high multiscale features from the dual encoders and up-sampling paths to recover the spatial resolution of the feature map at a fine granularity and deeper level. Finally, we directly expand the output of the Transformer branch to the same size as the ground truth to calculate the loss and simultaneously optimize the encoder and decoder in a multi-level optimization way, so that the network can further learn more semantic information and boundary details, and refine the segmentation results. Each component of SEAformer is detailed as follows.

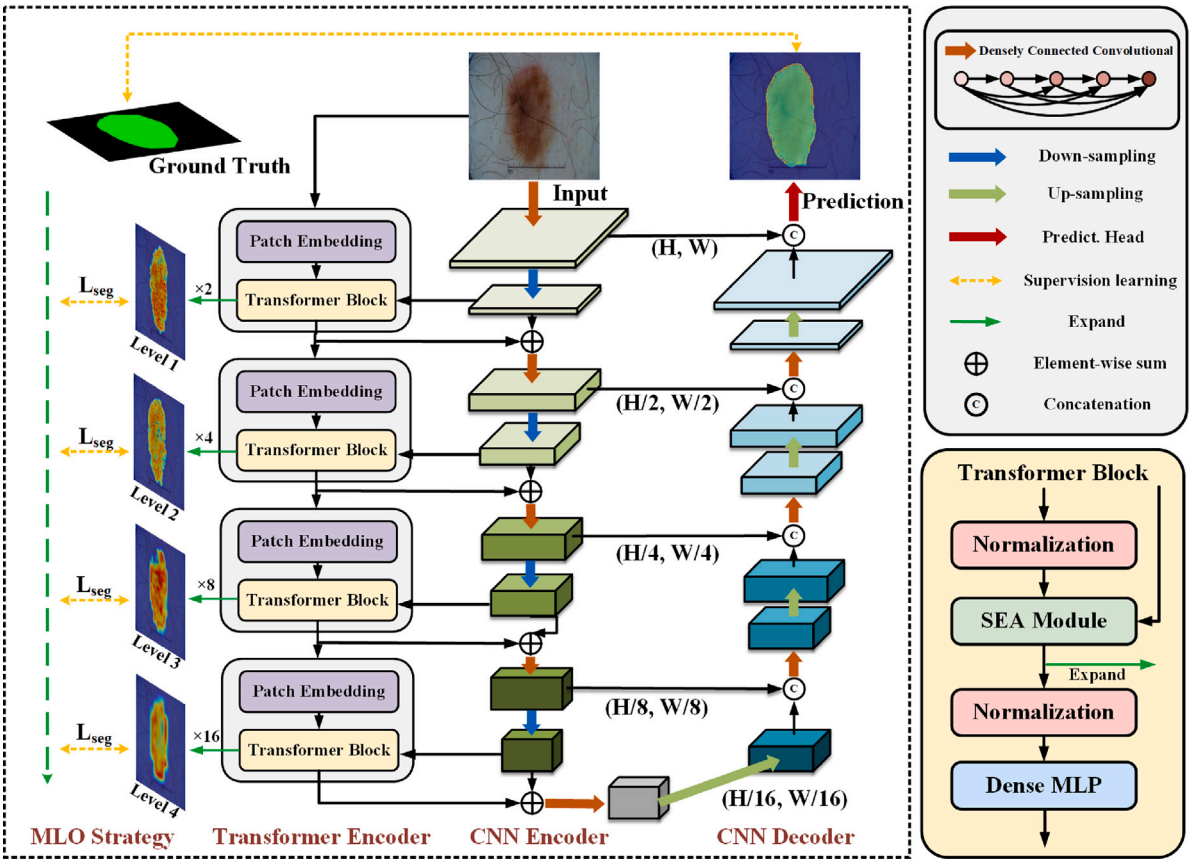

Fig. 2. Overview of the proposed SEAformer framework, which consists of a dual encoder with CNN and Transformer.

3.1. Transformer-based encoder

Currently, most Transformer-based medical image segmentation methods [11,22,32] use the standard Transformer [20] as an encoder. The standard Transformer consists of l identical blocks, each of which contains three components: the Normalization, the multi-head self-attention (MSA), and the Multi-layer Perceptron (MLP) layer. Residual connectivity is also applied after the MSA layer and the MLP layer, and the output of each block is expressed in Eq. (1):

$$\begin{aligned} \hat{X}^l &= MSA\left(Norm\left(X^{l-1}\right)\right) + X^{l-1} \\ X^l &= MLP\left(Norm\left(\hat{X}^l\right)\right) + \hat{X}^l \end{aligned} \tag{1}$$

In the standard Transformer architecture, the image is divided into patches of fixed size (16 × 16).

Inspired by methods such as [20,21,34], we redesign the Transformer structure. As shown in Fig. 2, we perform patch embedding on the input features before executing the Transformer block to reduce the resolution of the input feature map, so that the Transformer can obtain the features with pyramidal distribution and expand the receptive field layer by layer like the CNN. Specifically, the input feature map $X_{in} \in R^{H\times W\times C}$ is first sampled pixel by pixel, and the W and H information is concentrated in the channel dimension so that the channel is expanded to four times the original size, i.e. $\hat{X} \in R^{\frac{H}{2}\times\frac{W}{2}\times 4C}$. Then, the channels of $\hat{X}$ are mapped to the same channel dimension $X_{emb} \in R^{\frac{H}{2}\times\frac{W}{2}\times C}$ as the input feature map using a grouped convolution with a convolution kernel of 1 and a grouping number of 4. The Transformer block consists of three components: Normalization layer, SEA module, and Dense MLP layer. The output for each component is:

$$\begin{aligned} \hat{X}^l &= SEA\left(Norm\left(X_{emb}^{l-1}\right), X_{cnn}^{l-1}\right) \\ X^l &= DenseMLP\left(Norm\left(\hat{X}^l\right)\right) \end{aligned} \tag{2}$$

where X_{emb}^{l-1}, X_{cnn}^{l-1} denote the features from the Transformer branch after patch embedding and the features from the CNN branch, respectively. In training SEAformer, $\hat{X}^l$ is used to optimize the encoder. Next, we elaborate on the two components of the SEA module and Dense MLP.

3.1.1. SEA module

Much of the current work places a strong emphasis on attention and designing various attention-based components. In contrast, these works pay little attention to the role of the Transformer's overall architecture. We build on the PoolFormer by replacing the MSA in the standard Transformer block with the SEA module to focus on target edge information.

As shown in Fig. 3(a), the SEA module receives features from both Transformer and CNN branches. Since CNN can better capture the spatial information of the segmentation target, we supplement the Transformer branch with the CNN branch so that the two branches achieve feature fusion and complement each other.

CNN branch. In this branch, we use two very useful components, namely edge extraction block (EEB) and salient feature selection (SFS) (see Fig. 3(b) and (c)). These two components can extract information related to the target boundary and perform salient feature selection.

For SFS, we first aggregate the channel information of X_{sig} by depth, i.e., $X_{agg} = \sum_{c=1}^{C} X_{sig}$. Based on visual observation, the higher the activation response value in X_{agg}, the more likely it is to correspond to the target object. Then, we calculate the average value of all positions in X_{agg} as the threshold to select saliency features [35]:

$$X_{SFS}^{(i,j)} = \begin{cases} 1 & \text{if } X_{agg}^{(i,j)} > \bar{X}_{agg} \\ 0 & \text{other} \end{cases} \tag{3}$$

Here (i, j) denotes the coordinates of a particular location, $i, j \in [0, 1, \ldots, H-1], [0, 1, \ldots, W-1]$, and $X_{SFS} \in R^{H\times W\times 1}$.

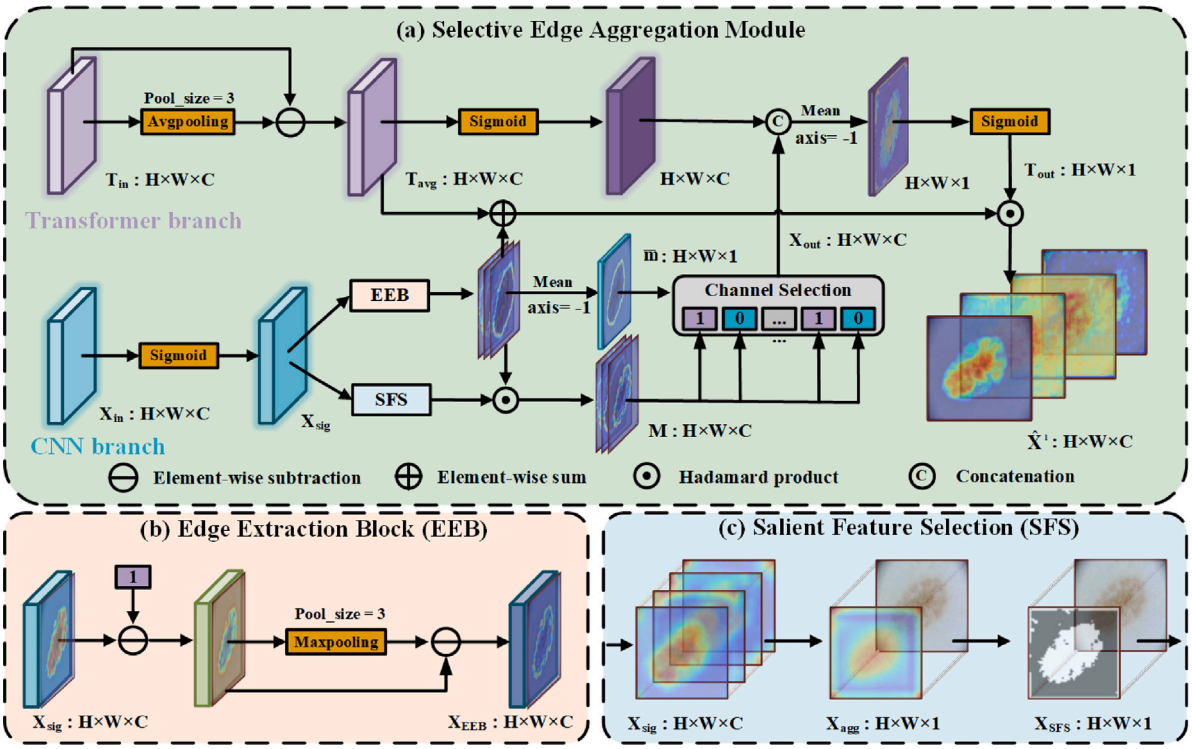

Fig. 3. (a) Selective edge aggregation module, (b) Edge extraction block, (c) Salient feature selection.

It can be observed from Fig. 3(c) that the selected features usually shield most of the noise, which is very useful for the model to locate the target region. To further obtain a feature map $M \in R^{H\times W\times C}$ that both masks the background region and extracts the target boundary, we multiply X_{SFS} and X_{EEB} element-by-element. However, in practical applications, we find that not all feature maps can meet our needs (feature maps with little noise background and distinct boundary demarcation). For this reason, we design a simple but effective algorithm to keep the feature maps that meet the expected effect as the output of the CNN branch and mask out the feature maps that do not meet the expectation. This algorithm uses the information of the convolution activation map itself and can be implemented by simply traversing each channel. The channel selection algorithm is described in Algorithm. 1 of the supplementary material.

Transformer branch. We utilize the intermediate and final outputs of the CNN branch for multi-level feature fusion with the Transformer branch so that the Transformer branch focuses on boundary information while obtaining more isomorphic features required for semantic segmentation.

First, we use averaging pooling with a window size of 3 to aggregate features near each patch of the input feature T_{in} and make each pixel of the output feature T_{avg} associated with its neighboring K^2 pixels:

$$T_{avg} = Avgpooling(T_{in}, K) - T_{in} \tag{4}$$

Then, the output of the Transformer branch is obtained by the following three steps: (1) feature concatenation of the activated T_{avg} with the output X_{out} of the CNN branch; (2) average all channel information of (1); (3) using the *Sigmoid* activation function (denoted by f in Eq. (5)) on the aggregated feature map to obtain the weight map. In summary, the above three steps can be expressed as:

$$T_{out} = f\left[\frac{1}{C}\sum_{c=1}^{C}\left(f\left(T_{avg}\right)©X_{out}\right)\right] \tag{5}$$

Finally, we add the boundary weight values to X_{EEB} and T_{avg} to increase the boundary weight value and multiply T_{out} as a weight map with the above results element by element to obtain the output of the SEA module, SEA_{out} is equivalent to $\hat{X}^l$ in Eq. (2).

$$SEA_{\text{out}} = T_{out} \odot \left(X_{EEB} + T_{avg}\right) \tag{6}$$

3.1.2. Dense MLP

Each layer in our Transformer block contains a series of densely connected feed-forward networks. Let the output feature map of the SEA module be $\hat{X}_0 \in R^{H\times W\times C}$. To accommodate the input requirements of the linear layer, we first reshape $\hat{X}_0$ into $\hat{X}_0 \in R^{N\times C}$, where $N = HW$. Then all subsequent layers of Dense MLP are related to all previous layers, i.e., $\hat{X}_0, \hat{X}_1, \ldots, \hat{X}_{k-1}$ are used as inputs for the next layer:

$$\begin{aligned}
&\hat{X}_1 = MLP(MLP(\hat{X}_0, 4*G), G)©\hat{X}_0,\\
&\hat{X}_2 = MLP(MLP(\hat{X}_1, 4*G), G)©\hat{X}_1,\\
&\ldots\ldots,\\
&\hat{X}_k = MLP(MLP(\hat{X}_{k-1}, 4*G), G)©\hat{X}_{k-1}
\end{aligned} \tag{7}$$

where G denotes the growth rate of the channel, which is the output dimension of the MLP layer ($G = 32$ in this paper). The Normalization layer and activation layer after the MLP layer are ignored in the above equation.

It can be seen that Dense MLP is different from vanilla MLP and ResMLP [36] in that it adds G channels sequentially to the output features of the previous layer, and obtains the output feature map $\hat{X}_k \in R^{N\times 2C}$ after k-1 times of dense connectivity (Due to the reduced spatial resolution, the number of output channels is increased to twice the input to prevent feature information loss). Finally, we reshape $\hat{X}_k$ to recover its spatial resolution to obtain the output $X \in R^{H\times W\times 2C}$ of Dense MLP, which achieves feature reuse with connections at the channel level.

3.1.3. Our transformer encoder vs. modern vision transformer architectures

The Transformer encoder we design is related to the modern vision Transformer methods [20,21,34], but is also quite different from them:

(1) Introduction of a completely non-parametric SEA module to replace the MSA module to improve the accuracy and interpretability of medical image segmentation by focusing on target edge information;

(2) Using Patch Embedding to make the feature maps of different stages into a feature pyramid distribution, preserving a more complete pixel space for the input features;

(3) The features of the CNN branch are introduced to compensate for the disadvantage of Transformer in mining spatial information, which enables the model to converge quickly even when training small datasets;

(4) The first attempts to construct a dense MLP by applying linear mapping in the channel direction using dense connections, which enables the model to achieve better performance with smaller parameters.

3.2. Densely connected CNN-based encoder and decoder

As shown in Fig. 2, the CNN-based encoder and decoder is a U-shaped network, where the encoder is used to extract semantic information from shallow to deep in medical images, and the decoder is used to recover the spatial resolution of the encoder output features. In addition, a skip connection is applied to obtain detailed information from the encoder and decoder to compensate for the loss of information due to down-sampling and convolution operations. Considering the need to fully utilize the global context information extracted by the Transformer encoder branch, we down-sample twice in order from the first block, and the final resolution becomes $(H/16, W/16)$ so that we can not only obtain rich local information but also perform flexible feature fusion with the Transformer branch at the same depth. If X_{cnn}^{l-1} and X_{trans}^{l-1} are used to denote the output of the $(l-1)$st, block of the CNN encoder and Transformer encoder, respectively, then the fusion feature used for the skip connection can be expressed as:

$$X_{con}^{l} = DenseConv(X_{cnn}^{l-1} \oplus X_{trans}^{l-1}) \tag{8}$$

where "$\oplus$" denotes element-wise summation and DenseConv represents a densely connected convolutional block. In practical applications, we can build the DenseConv block ourselves, or we can directly use the pre-trained model of DenseNet121 [37] (or another backbone).

To obtain accurate pixel-level prediction results, we also design a decoder based on a densely connected CNN. Because the feature cascade of X_{con}^{l} and decoder increases the number of channels to twice the original number, if the subsequent operation is performed directly, it will not only increases the operational load of the network but also causes feature redundancy. To balance efficiency and reduce redundant features, we use a standard convolution to reduce the cascaded channels to $1/4$ of the original and then increase the number of channels to $1/2$ of the original channels through a series of densely connected convolutions. Compared with conventional decoders that reduce the channels to $1/2$ of the original directly after each layer cascade, SEAformer deepens the network depth without increasing the network load and allows full information integration at different scales. Finally, the prediction maps are generated by a 1×1 convolution layer with a $Sigmoid$ activation function.

3.3. Training and inference details

Loss function. To reduce the difference between the prediction result (P) and ground truth (G), two loss functions are used in this paper to focus on two separate aspects of the segmentation content and segmentation boundary, respectively [38]. The first one is the IoU loss, which is used to minimize the overlap error between P and G. A large number of previous practices have proved the feasibility of IoU loss in different medical image segmentation tasks, which can be expressed as:

$$\ell_{IoU} = -\log \frac{\sum_i P_i G_i}{\sum_i P_i + \sum_i G_i - \sum_i P_i G_i} \tag{9}$$

where i denotes the index of all pixels in G and P. Unlike the previous method that uses $1 - IoU$ as the loss function, we use $-\log(IoU)$ to calculate the loss, which can make the network converge more smoothly.

The second one is the edge loss, which is used to minimize the boundary error between P and G. Similar to the SEA module, we first extract the boundaries of P and G using the pooling operation, which can be expressed as:

$$\begin{aligned} G^b &= Maxpool(1-G, K) - (1-G), \\ P^b &= Maxpool(1-P, K) - (1-P) \end{aligned} \tag{10}$$

where $K = 3$, which denotes the sliding window size for the max-pooling. Then, we construct the edge loss using G^b and P^b:

$$\ell_{\text{Edge}} = \frac{\sum_i P_i^b G_i^b}{\sum_i P_i^b G_i^b + \alpha \sum_i (1-P_i^b) G_i^b + (1-\alpha) \sum_i P_i^b (1-G_i^b)} \tag{11}$$

where G_i^b and P_i^b denote the ground truth and predicted boundary probability values at i locations, respectively. Since we highlight the information of boundary pixels, the number of lesion pixels will be much lower than the number of non-lesion pixels (only boundary pixels are represented as lesion pixels in G^b). We trade off the high imbalance in the number of pixels by setting $\alpha = 0.75$ to make the network more concerned with the recall of the boundaries.

Finally, our loss function consists of ℓ_{IoU} and ℓ_{Edge}:

$$\ell_{Seg} = \lambda_1 \ell_{IoU} + \lambda_2 \ell_{Edge} \tag{12}$$

Empirically, we set $\lambda_1 = 0.6$, $\lambda_2 = 0.4$ respectively.

Training and Inference. As shown in Fig. 2, to obtain high-quality mask and boundary information during the training phase, we perform multi-level optimization (MLO) on the output probability map P_d of the decoder and the output P_e of the encoder. Note that the output feature maps of the encoder are up-sampled to the same size as the ground truth when supervised learning is performed (i.e., up-sampled ×2, ×4, ×8, ×16 times, respectively). Therefore, the overall loss of the training phase can be written as:

$$\ell_{Total} = \ell_{Seg}\left(G, P_d\right) + \frac{1}{N} \sum_{i=1}^{N} \ell_{Seg}\left(G, P_e^i\right) \tag{13}$$

In the inference phase, we can directly obtain the output probability map P_d of the decoder as the final output result. This design infers significant object regions through a top-down workflow, effectively guiding the network to explore the details of the object and optimize the boundaries for more complete and explicit boundary prediction.

4. Experiments and results

4.1. Implementation details

We implement the Keras-based approach by training on an NVIDIA RTX3090 GPU (24 g). Using the Adam optimizer, the learning rate is fixed at 1e−4. The mini-batch size is set as 16, and the training is stopped using an early stop mechanism when the validation loss is stable and there is no significant change within 30 epochs. We expand the training data by applying random rotation (±25°), random horizontal and vertical shifts (15%), and random flips (horizontal and vertical). The same training and validation sets are used for all comparison experiments. After the second stage of the CNN branch, the initial weights come from the Block2, Block3, and Block4 of DenseNet121 pre-trained on ImageNet,[1] and the other layers we train from scratch.

To comprehensively and objectively evaluate the segmentation performance of the proposed method, we use five widely used metrics to evaluate the performance of SEAformer. Namely, Accuracy (Acc),

[1] https://github.com/liuzhuang13/DenseNet?tab=readme-ov-file.

Table 1
Comparison of different methods on the ISIC2017 and the PH2 dataset.

Method	ISIC2017 dataset					PH2 dataset					Params ↓	GFLOPs ↓	FPS ↑
	Acc ↑	Sens ↑	Spec ↑	IoU ↑	Dice ↑	Acc ↑	Sens ↑	Spec ↑	IoU ↑	Dice ↑			
U-Net [3]	0.925	0.823	0.974	0.746	0.834	0.923	0.913	0.956	0.841	0.894	30 M	42	**125**
U-Net++ [7]	0.928	0.797	0.981	0.734	0.823	0.951	0.945	0.941	0.863	0.921	25 M	29	77
R2U-Net [6]	0.934	0.837	**0.982**	0.767	0.851	0.946	0.953	0.935	0.855	0.916	92 M	103	46
DeepLabv3+ [39]	0.931	0.821	0.977	0.752	0.838	0.940	0.956	0.932	0.858	0.920	39 M	10	53
BiONet [8]	0.935	0.854	0.968	0.774	0.854	0.943	0.924	0.956	0.850	0.914	57 M	71	40
TransU-Net [11]	0.939	0.865	0.965	0.785	0.865	0.949	0.974	0.916	0.864	0.921	100 M	25	44
SwinU-Net [40]	0.936	0.861	0.964	0.775	0.856	0.952	**0.977**	0.921	0.856	0.918	26 M	**6**	57
ResGANet [41]	0.936	0.842	0.950	0.764	0.862	0.948	0.958	0.928	0.868	0.923	39 M	65	13
FATNet [22]	0.935	0.873	0.949	0.772	0.854	**0.961**	0.957	**0.963**	**0.871**	0.927	27 M	43	32
EANet [19]	0.936	0.847	0.967	0.773	0.854	0.946	0.974	0.918	0.852	0.914	30 M	99	26
H2Former [30]	0.934	0.846	0.968	0.756	0.840	0.952	0.959	0.929	0.866	0.923	33 M	34	36
HTC-Net [29]	0.939	**0.880**	0.951	0.779	0.861	0.957	0.970	0.947	0.867	0.924	56 M	13	26
SegNetr [31]	0.937	0.845	0.959	0.775	0.856	0.950	**0.977**	0.932	0.855	0.916	**12 M**	10	22
MSDANet [26]	0.934	0.825	0.972	0.757	0.841	0.949	0.962	0.911	0.856	0.918	52 M	124	19
SEAformer	**0.940**	**0.880**	0.963	**0.790**	**0.869**	0.957	**0.977**	0.944	0.870	**0.933**	16 M	32	26

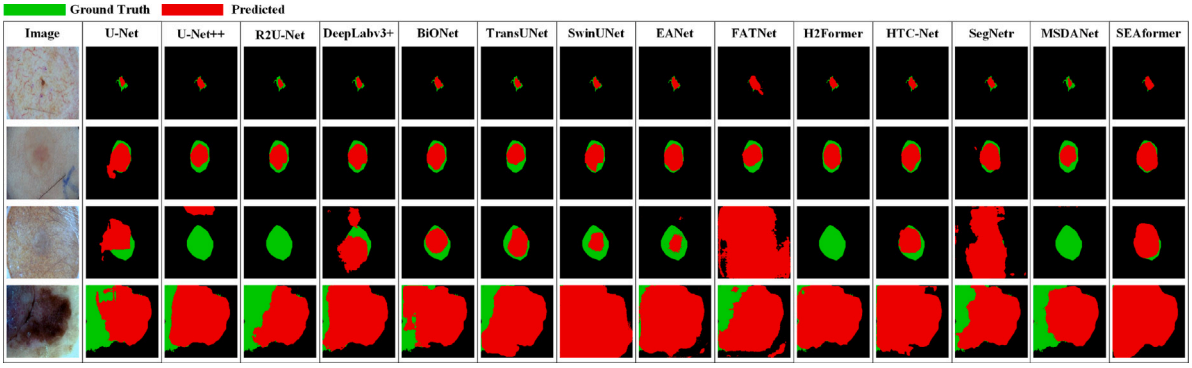

Fig. 4. Qualitative comparison results on skin lesion segmentation.

Sensitivity (Sens), Specificity (Spec), Intersection over Union (IoU), and Dice similarity coefficient (Dice). We also report the number of parameters, GFLOPs, and FPS of all compared methods for combined comparison.

4.2. Skin lesion segmentation

4.2.1. Dataset details

We performed experiments on two publicly available skin lesion segmentation datasets (ISIC2017 [42] and PH2 [43]). The ISIC2017 dataset provided by the International Skin Imaging Collaboration (ISIC), consists of 2000 training images, 150 validation images, and 600 test images. The PH2 dataset includes 200 skin mirror images with a resolution of 765 × 572 pixels. As in the literature [22], we randomly select 140 images as the training set, 20 images as the validation set, and the remaining 40 images as the test set. We first normalize the colors of the images using the grayscale world color consistency algorithm for the two datasets mentioned above, and then all the images are adjusted to 224 × 224 pixels resolution for the experiments. Finally, the training data is augmented to improve the generalization ability of the model during the training process.

4.2.2. Results

To compare the performance of skin lesion segmentation, we conducted quantitative and qualitative experiments on fourteen state-of-the-art methods. Among the CNN-based methods are U-Net [3], U-Net++ [7], R2U-Net [6], DeepLabv3+ [42], BiONet [8], ResGANet [44], and EANet [19]. CNN-Transformer-based methods include TransU-Net [11], SwinU-Net [43], FATNet [22], H2Former [30], HTC-Net [29], SegNetr [31], and MSDANet [26]. The code for these methods is provided open-source by the authors, and we executed them in the recommended environment to obtain experimental results.

The quantitative results for the ISIC2017 and PH2 datasets are presented in Table 1. SEAformer, with a parameter count of only 16 million, holds a significant advantage over other methods. On the ISIC2017 dataset, SEAformer notably surpasses the benchmark U-Net across all evaluation metrics. When compared to SwinU-Net, a model with a pure Transformer architecture, SEAformer exhibits improvements of 1.5% in IoU and 1.3% in Dice coefficient. Furthermore, SEAformer outperforms TransU-Net and FATNet, two methods that also leverage the strengths of both Transformer and CNN architectures. For the PH2 dataset, our approach achieves the highest Dice score. In comparison to EANet, a method focusing on boundary quality, SEAformer shows enhancements of 1.1% in accuracy and 1.8% in IoU. These results underscore the superiority of our proposed methodology. Notably, when contrasted with SegNetr, a model of similar parameter capacity, SEAformer achieves over a 1% improvement in IoU on both datasets while also demonstrating superior frames per second (FPS) performance. Noteworthy observations include the trend that methods combining CNN and Transformer advantages tend to yield superior segmentation outcomes compared to singular-structure approaches. Similarly, concentrating on boundary quality enhancement can elevate the performance of benchmark techniques.

Moreover, qualitative experimental results from the ISIC2017 dataset, as depicted in Fig. 4, underscore SEAformer's superiority over other leading segmentation approaches. Specifically, SEAformer's segmentation outputs on images with diverse lesion sizes and indistinct boundaries closely align with ground truth annotations.

Table 2
Comparison of different methods on the TN-SCUI dataset and the GLAS segmentation dataset.

Method	TN-SCUI dataset					GLAS segmentation dataset					Params ↓	GFLOPs ↓	FPS ↑
	Acc ↑	Sens ↑	Spec ↑	IoU ↑	Dice ↑	Acc ↑	Sens ↑	Spec ↑	IoU ↑	Dice ↑			
U-Net [3]	0.978	0.828	0.989	0.711	0.799	0.869	0.863	0.873	0.773	0.864	30 M	42	**125**
U-Net++ [7]	0.979	0.849	0.989	0.723	0.813	0.876	0.873	0.879	0.785	0.873	25 M	29	77
R2U-Net [6]	0.977	0.761	0.990	0.667	0.762	0.891	0.864	**0.910**	0.800	0.883	92 M	103	46
DeepLabv3+ [39]	0.980	0.855	0.989	0.731	0.824	0.872	0.866	0.879	0.777	0.866	39 M	10	53
BiONet [8]	0.977	0.839	0.987	0.713	0.803	0.870	0.827	0.903	0.758	0.852	57 M	71	40
TransU-Net [11]	0.981	0.844	0.990	0.737	0.826	0.887	0.882	0.898	0.801	0.884	100 M	25	44
SwinU-Net [40]	0.980	0.841	**0.991**	0.736	0.825	0.895	0.894	0.867	0.818	0.895	26 M	**6**	57
ResGANet [41]	0.978	0.830	0.990	0.726	0.818	0.882	0.878	0.883	0.792	0.878	39 M	65	13
FATNet [22]	0.981	0.856	0.989	0.742	0.830	0.887	0.893	0.871	0.800	0.884	27 M	43	32
EANet [19]	0.979	0.851	0.989	0.739	0.823	0.885	0.892	0.864	0.794	0.880	30 M	99	26
H2Former [30]	0.982	0.853	0.990	0.749	0.839	0.898	**0.900**	0.883	0.817	0.894	33 M	34	36
HTC-Net [29]	0.982	**0.865**	0.989	0.753	0.840	0.889	0.895	0.848	0.809	0.890	56 M	13	26
SegNetr [31]	0.980	0.853	0.988	0.745	0.831	0.896	0.897	0.881	0.819	0.890	**12 M**	10	22
MSDANet [26]	**0.983**	0.852	0.990	0.757	0.843	0.869	0.882	0.846	0.773	0.866	52 M	124	19
SEAformer	**0.983**	0.852	**0.991**	**0.767**	**0.846**	**0.904**	0.897	0.900	**0.826**	**0.901**	16 M	32	26

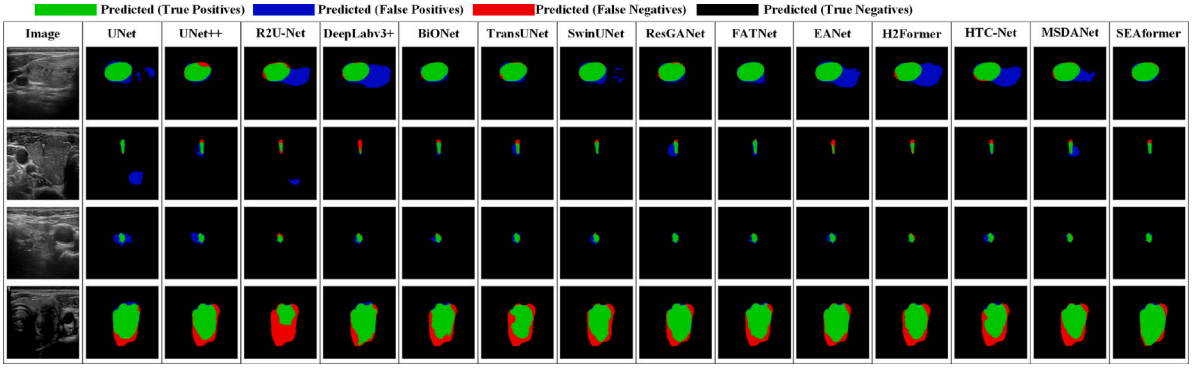

Fig. 5. Qualitative comparison results on thyroid nodules segmentation.

4.3. Thyroid nodules segmentation

4.3.1. Dataset details

This dataset is part of the TN-SCUI 2020 challenge.[2] All personal labels of the scanned images is removed to protect the privacy of the patients. The dataset provides 3644 nodular thyroid images of different sizes and the nodules have been annotated by experienced physicians. In our experiments, the dataset is divided into a training set (60%), a validation set (20%), and a testing set (20%). The enhancement strategy described in Section 4.1 is used to increase the diversity of the training data during the training process, and we uniformly adjust the resolution of all images to 224 × 224.

4.3.2. Results

As shown in Table 2, we conducted a quantitative comparison of SEAformer against fourteen state-of-the-art segmentation models. Compared to skin lesion segmentation tasks, thyroid nodule segmentation is more challenging due to the relatively blurred boundaries. SEAformer demonstrated excellent performance in both Dice and IoU metrics, achieving 0.846 and 0.767, respectively, surpassing most competing methods. The high Dice score indicates that SEAformer provides more precise predictions for segmentation boundaries and target regions, while the high IoU confirms better overlap in the overall segmented areas. In terms of parameter count (Params) and computational cost (GFLOPs), SEAformer uses only 16M parameters and 32 GFLOPs, significantly lower than most existing models. For instance, TransU-Net and R2U-Net have 100M and 92M parameters, respectively, while SEAformer requires only about one-sixth of these. This highlights the efficiency of SEAformer's model design, which reduces parameter count and computational overhead through innovative architectural approaches. These experimental results demonstrate that SEAformer effectively addresses the limitations of Transformer-based models in capturing fine details and boundary information through its improved SEA module and multi-level optimization strategies. Even for tasks with blurred boundaries, SEAformer achieves superior segmentation performance.

Fig. 5 further illustrates the qualitative results of different methods. It can be observed that SEAformer shows a clear advantage in accurately delineating nodule boundaries. Across all samples, the green areas (true positives) generated by SEAformer closely align with the ground truth, indicating precise segmentation without noticeable over-segmentation (red areas) or under-segmentation (blue areas). This boundary precision is particularly crucial for thyroid nodule segmentation tasks, where edges can be blurred and difficult to distinguish. The SEA module in SEAformer appears to effectively enhance edge-related features, allowing for more accurate boundary predictions compared to models like U-Net and TransU-Net, which tend to struggle with more irregular shapes.

4.4. GLAnd segmentation

4.4.1. Dataset details

The GLAnd segmentation (GLAS) dataset [45] contains microscopic images of hematoxylin and eosin-stained slides, as well as ground truth provided by expert pathologists. The dataset contains 165 images that are not uniform in resolution size, with a minimum resolution of 433 × 574 and a maximum resolution of 775 × 522. We use 85 images for training and 80 images for testing. The resolution of all images in the experiments is adjusted to 224 × 224.

[2] https://tn-scui2020.grand-challenge.org/.

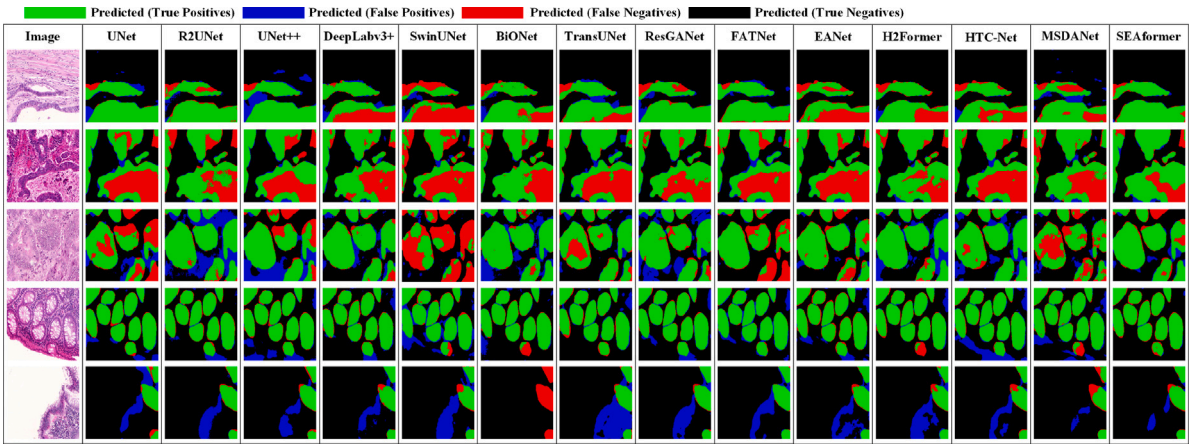

Fig. 6. Qualitative comparison results on GLAnd segmentation.

Table 3
Comparison of different methods on the COVID-19 infection segmentation dataset.

Method	Acc% ↑	Sens% ↑	Spec% ↑	IoU% ↑	Dice% ↑
U-Net [3]	97.22 ± 0.35	77.48 ± 4.02	98.34 ± 0.21	61.98 ± 3.50	74.98 ± 3.31
U-Net++ [7]	96.44 ± 0.49	65.29 ± 3.41	98.30 ± 0.37	52.79 ± 2.30	66.95 ± 1.94
R2U-Net [6]	96.64 ± 0.33	75.26 ± 3.32	98.22 ± 0.19	60.79 ± 2.58	74.68 ± 3.25
DeepLabv3+ [39]	97.00 ± 0.25	76.36 ± 3.36	98.52 ± 0.21	61.82 ± 2.56	76.08 ± 3.42
BiONet [8]	96.76 ± 0.35	74.24 ± 3.36	98.06 ± 0.39	59.16 ± 2.80	73.12 ± 2.47
TransU-Net [11]	97.54 ± 0.21	79.54 ± 4.10	**98.64 ± 0.24**	65.50 ± 3.44	78.32 ± 2.60
SwinU-Net [40]	97.03 ± 0.49	77.09 ± 3.47	98.06 ± 0.55	61.24 ± 2.05	74.63 ± 1.64
ResGANet [41]	96.92 ± 0.44	77.62 ± 3.07	98.44 ± 0.27	64.06 ± 2.67	76.97 ± 2.08
FATNet [22]	97.50 ± 0.27	77.32 ± 3.36	**98.64 ± 0.27**	64.76 ± 3.17	77.82 ± 2.38
EANet [19]	97.32 ± 0.29	76.98 ± 2.62	98.46 ± 0.27	63.84 ± 2.81	76.84 ± 2.39
H2Former [30]	96.66 ± 0.55	74.22 ± 3.59	97.80 ± 0.45	58.62 ± 1.93	72.18 ± 1.83
HTC-Net [29]	97.02 ± 0.36	76.65 ± 2.63	98.10 ± 0.32	60.59 ± 1.95	74.22 ± 1.33
SegNetr [31]	97.52 ± 0.15	78.28 ± 3.32	98.04 ± 0.22	64.92 ± 3.40	77.28 ± 3.03
MSDANet [26]	96.12 ± 0.40	75.95 ± 2.44	98.49 ± 0.29	61.59 ± 1.99	75.24 ± 1.53
SEAformer	**97.68 ± 0.26**	**80.26 ± 3.10**	**98.64 ± 0.18**	**66.00 ± 3.47**	**78.68 ± 2.50**

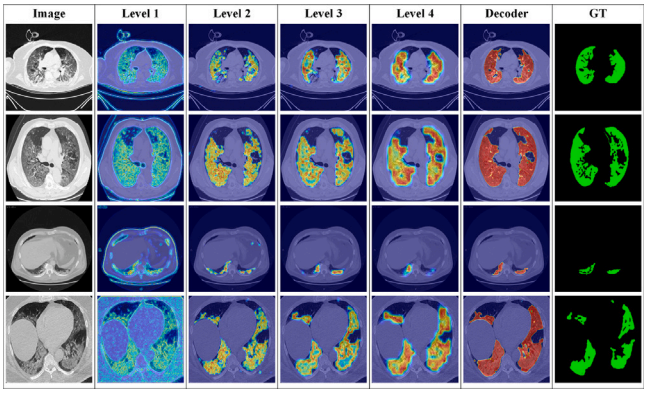

Fig. 7. Visual comparison of each encoder (level1–level4) and decoder block feature mapping.

4.4.2. Results

Table 2 presents the quantitative experimental results of SEAformer compared with other state-of-the-art methods on the GLAS segmentation dataset. The proposed method achieves the highest values in three key metrics — Accuracy, IoU, and Dice (i.e., 0.904, 0.826, and 0.901, respectively) — surpassing FATNet, which also employs a dual-encoder structure. Compared to TransU-Net, which incorporates a Transformer architecture, SEAformer demonstrates improvements across five metrics, with a particularly notable increase of 2.5% in the IoU metric. SEAformer achieves these high performance levels with a relatively low number of parameters (16M) and GFLOPs (32), making it much more efficient than most other models. This lightweight design is particularly advantageous for real-world applications with limited computational resources. Although SEAformer is not the fastest model, its speed is competitive and suitable for real-time applications. In addition to the quantitative results, we also provide qualitative comparisons between SEAformer and nine other state-of-the-art methods, as shown in Fig. 6. Visually, SEAformer's predictions are closer to the ground truth, especially with lower false positives (i.e., blue areas), which is crucial in clinical applications. SEAformer also demonstrates finer detail and higher accuracy along the segmentation boundaries than other methods, further underscoring its effectiveness for precise medical image segmentation.

4.5. COVID-19 infection segmentation

4.5.1. Dataset details

The COVID-19 Infection segmentation dataset[3] contains 100 axial CT images and corresponding annotated images from more than 40 COVID-19 patients. Considering the very small amount of data in this dataset, we conduct experiments using five-fold cross-validation (i.e., using 80 images for training and 20 images for validation each time). For training, we also use a data enhancement strategy to increase the diversity of the training set and adjust the images to a uniform 352 × 352 resolution.

[3] https://medicalsegmentation.com/covid19/.

Table 4
Ablation analysis of different components of SEAformer is performed on ISIC2017, PH2, TN-SCUI and GLAS datasets.

ID	Baseline	Pretrain	Trans	SEA	MLO	ISIC2017		PH2		TN-SCUI		GLAS		#Param	FLOPs	FPS
						IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice			
(a)	✓					0.769	0.850	0.843	0.900	0.750	0.837	0.811	0.892	14 M	31 G	33
(b)	✓	✓				0.777	0.858	0.865	0.921	0.756	0.843	0.815	0.895	14 M	31 G	39
(c)	✓	✓	✓			0.781	0.861	0.865	0.921	0.763	0.850	0.814	0.895	16 M	32 G	28
(d)	✓	✓	✓	✓		0.784	0.863	0.847	0.918	0.767	0.852	0.820	0.899	16 M	32 G	26
(e)	✓	✓	✓	✓	✓	0.790	0.869	0.870	0.933	0.771	0.855	0.826	0.901	16 M	32 G	26

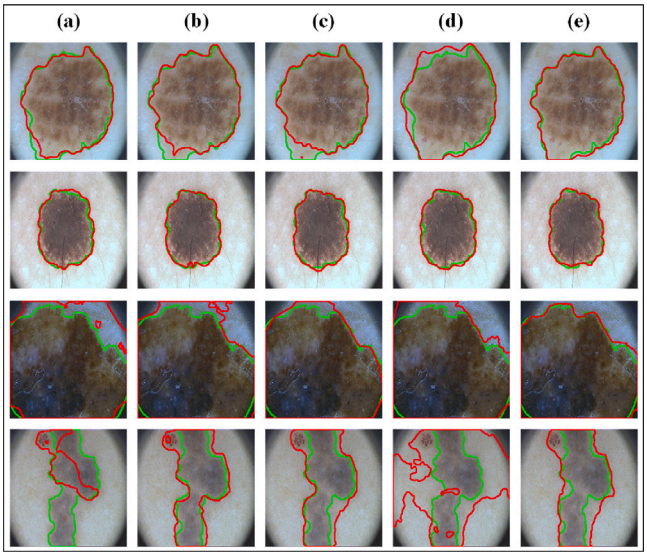

Fig. 8. Qualitative experimental results of different components of SEAformer on the PH2 dataset.

4.5.2. Results

To evaluate the performance of different methods on a very small-scale training dataset, we used the COVID-19 infection segmentation dataset, where the lesions are smaller and more dispersed compared to other datasets. We compared SEAformer with 14 other methods, and Table 3 shows the results from a five-fold cross-validation.

By incorporating a Transformer to capture long-range dependencies, models like TransU-Net and FTANet achieved better segmentation performance than U-Net. Although R2U-Net and BiONet introduced recurrent structures into U-Net, they did not outperform U-Net on this dataset, suggesting that recurrent structures may be less suitable for small datasets. Interestingly, we found that dual-encoder networks generally achieved better results than single-encoder networks. For example, FATNet attained an average IoU of 64.76%. SEAformer also benefited from the dual-encoder structure, achieving the best average IoU (66.00%), Dice score (78.68%), and Sensitivity (80.26%). We attribute the effectiveness of the dual-encoder structure to two main factors: first, it can leverage pre-trained models to capture deeper feature information, allowing for faster model convergence. Second, the encoder trained from scratch helps address domain shifts that arise from weight transfer and compensates for lost detail information. In addition to the quantitative results, we also performed a qualitative comparison on challenging cases within this dataset. Finally, we visualize the four stages of the SEAformer encoder and the feature mapping of the decoder at different levels on both COVID-19 dataset (as shown in Fig. 7). It can be seen that the shallow network (e.g., Level1 and Level2) focuses more on texture, grayscale, etc. As the depth becomes deeper, the network can learn deeper semantic information to compensate for the information loss problem of the shallow network and reduce the weight of interfering information, thus locating the target area more accurately. In addition, the visualization results of the decoder reveal that the segmentation boundaries are clear and weight more than the surrounding pixels (exhibiting higher thermal values), indicating that the training strategy in this paper helps to learn a better representation and produce more accurate masks.

4.6. Ablation studies

Effectiveness of SEAformer Components. Table 4 presents the ablation study on the different components of SEAformer. We use a densely connected CNN-based encoder and decoder (Section 3.2) as the baseline model (a), and SEAformer is finally constructed by adding the Transformer branch (c), the selective edge aggregation module (d), and the multi-level optimization strategy (e). Based on prior experience, networks with pretrained weights tend to be easier to optimize and converge faster than networks trained from scratch, as they can learn valuable feature information from large-scale datasets. We observe that models with pretrained (b) weights show various degrees of improvement across different datasets. Compared to the baseline (b), the dual-encoder approach with Transformer shows performance gains on the ISIC2017 and TN-SCUI datasets. However, no improvement is observed on the smaller PH2 and GLAS datasets, likely due to the limited benefits of the Transformer architecture on small datasets. Additionally, incorporating the Transformer encoder does not significantly increase the parameter count (14M vs. 16M) or GFLOPs (31 vs. 32). The results in the (d) validate the effectiveness of the SEA module. After adding the SEA module, the model's IoU and Dice scores on the ISIC2017 dataset improve by 0.3% and 0.2%, respectively. We also note that the performance gains from the SEA module are more pronounced on larger datasets, likely due to the richer spatial features learned by the CNN encoder. Finally, the MLO strategy (e) combines ℓ_{IoU} and ℓ_{Edge} losses, optimizing both the encoder and the entire network to enhance segmentation performance. As shown in the last row, the network trained with the MLO strategy effectively improves segmentation accuracy, demonstrating its effectiveness in boosting overall performance.

To further analyze why SEAformer did not achieve consistent performance improvements on the PH2 dataset, we conducted a visualization study (Fig. 8). As shown in the top three rows, when the lesions are relatively clear, even the simple baseline method can effectively localize the lesion areas. With the addition of pretraining (b) and the dual-encoder structure (c), the segmentation results become closer to the ground truth. However, upon introducing the SEA module, we observed no significant improvement in segmentation results for certain cases, such as those in the first and last rows. The primary error involves misclassifying background regions as lesion areas. We believe this misclassification occurs because the features of these regions resemble those of the lesions, causing the model to mistakenly identify them as lesion areas. Additionally, potential annotation errors cannot be ruled out in these images. Finally, the relatively small size of the PH2 dataset may have limited the model's learning capacity, leading to inaccuracies in the decision-making process.

Effectiveness of Growth Rate (G) in SEAformer. We conducted an ablation study on SEAformer with different growth rate configurations. In this setting, "G = 16/32" indicates that the growth rates for DenseConv and Dense MLP in SEAformer are set to 16 and 32, respectively. Fig. 9 displays the variations in IoU and Dice scores across four datasets. As observed, the "G = 16/32" configuration achieved optimal performance

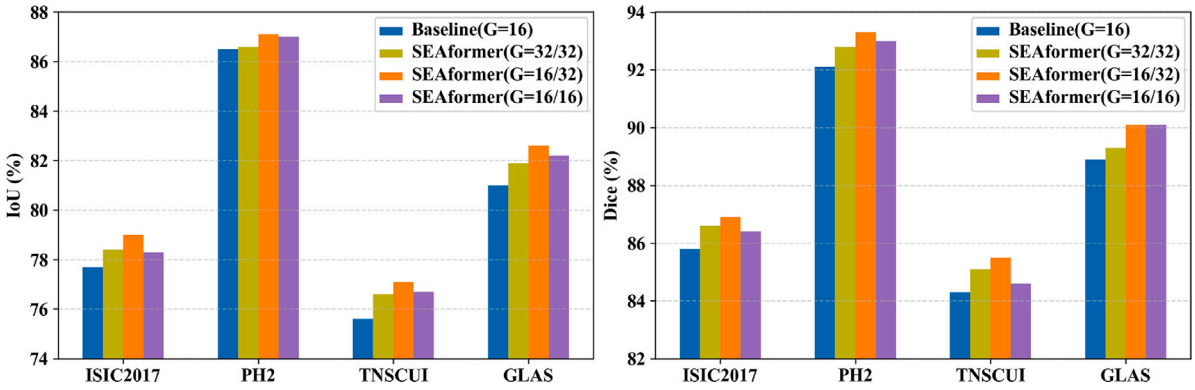

Fig. 9. Ablation study on growth rates of SEAformer across four datasets: IoU (left) and Dice (right).

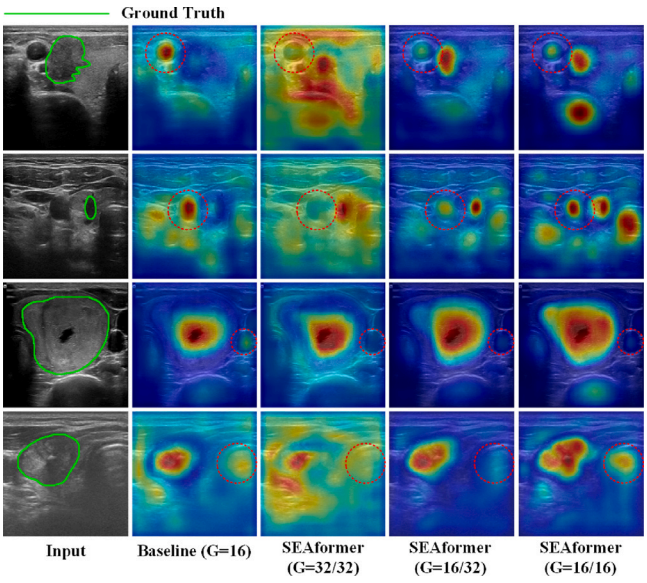

Fig. 10. Comparison of feature attention heatmaps for the highest stage of the SEAformer encoder with different configurations.

on most datasets. Additionally, Fig. 10 provides visualizations of the feature attention heatmaps from the top layer of the SEAformer encoder under different configurations on the TN-SCUI dataset. We found that the dual-encoder approach performed better in locating lesion areas compared to the single-encoder baseline, evidenced by higher weight values in the lesion regions. When the growth rate of DenseConv was set to 32, we did not use pre-trained parameters, resulting in slightly lower localization performance for SEAformer (G = 32/32) compared to other configurations. The red dashed circles highlight areas with features similar to those of lesions, where the model is prone to mislocalization. It is evident that the SEAformer with the "G = 16/32" configuration displays relatively lower weight values in non-lesion areas. We believe that this configuration enhances lesion localization capability primarily due to the following two points: firstly, the introduction of pre-trained models enables the model to more accurately locate target regions; secondly, the configuration with different growth rates allows the model to learn features at different scales, achieving complementarity among multi-scale features. Therefore, the SEAformer designed with different growth rates provides better boundary details and lesion localization capabilities.

Effectiveness of Loss Function. As shown in Table 5, we conducted a comparative study of various loss functions used for training SEAformer across four different datasets to assess the effectiveness of the proposed

Table 5
Performance comparison of SEAformer with different loss functions.

Method	ISIC2017		PH2		TN-SCUI		GLAS	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
BCE	0.776	0.858	0.852	0.917	0.749	0.830	0.812	0.891
Dice	0.782	0.863	0.870	0.932	0.758	0.839	0.816	0.893
IoU	0.780	0.862	0.866	0.922	0.758	0.840	0.813	0.893
BCE+IoU	0.785	0.866	0.872	0.935	0.766	0.853	0.824	0.892
IoU+Edge	0.790	0.869	0.870	0.933	0.771	0.855	0.826	0.901

IoU+Edge loss. Specifically, we compared our proposed loss function with commonly used alternatives, including Binary Cross-Entropy (BCE), Dice, IoU, and BCE+IoU losses. The results indicate that IoU and Dice losses perform better than BCE, as they focus more on the overlap ratio between the predicted results and ground truth, leading to improved segmentation accuracy. Multi-loss approaches provide complementary guidance to the model's optimization, enhancing its overall performance. Compared to the widely used BCE+IoU loss, our proposed IoU+Edge loss achieves superior segmentation performance across all datasets, underscoring its effectiveness in boosting segmentation accuracy. These findings confirm that the added edge-awareness component in the IoU+Edge loss significantly improves boundary precision and overall model performance.

5. Discussion and limitations

Although many encoder–decoder medical image segmentation models have been proposed so far, most of them only consider how to make their methods effective in a specific task, ignoring the problems of limited training data of medical images, the diversity of pathological images, the large variation of lesion scales and the blurred lesion boundaries. Most existing methods still cannot accurately and robustly classify target and background pixels due to limitations in mining global contextual dependencies and the inability to delineate target boundaries clearly. In this paper, to obtain a better feature representation, we integrate two branches of CNN and Transformer in the encoder to extract local and global information of medical images. Since we use dense connectivity throughout the network architecture to extract features, our SEAformer has a natural multiscale capability without additional complex multiscale modules. Furthermore, we believe that focusing on boundary information is a promising approach to improve the performance of medical image segmentation. We equip the Transformer block of SEAformer with a completely non-parametric selective edge aggregation module that locates the target boundaries and aggregates this information based on the strengths of the CNN branch in spatial feature extraction, enabling the network to better distinguish the target boundaries from the background. Unlike traditional optimization, we optimize each stage of the encoder and the decoder

simultaneously, which not only allows the codec to gradually absorb useful information and achieve complementarity but also effectively guides the network to explore the missing parts and details of the object, resulting in a more complete mask.

The results of our extensive ablation studies and comparative experiments on several medical image segmentation tasks have demonstrated their advantages, and the ideas presented in this paper will also provide some insights for researchers working on building encoder–decoder approaches for medical image segmentation tasks. Firstly, medical images have limited training data, and although using pre-training weights can help models fit the data faster and achieve higher accuracy, researchers should use pre-training weights flexibly according to the specific task and model architecture, rather than directly applying the existing backbone. Secondly, the interaction of local and global contextual information is very important, and the model can learn complementary information in the interaction. Also, the boundary information can be considered complementary information to the segmentation mask. Thirdly, choosing the appropriate optimization strategy can help the model to obtain more accurate and robust experimental results. Finally, we believe that approaches that mix the advantages of CNN and Transformer are more effective than those using any single structure. More and more similar methods will be proposed and used for various challenging medical computer vision tasks in the future, while our model provides some initial experience in designing powerful hybrid approaches.

Although our proposed method has shown promising results, it has some limitations that need to be addressed in future studies: (1) In the experiments of this paper we can see that the encoder can produce slightly worse segmentation results than the decoder by direct up-sampling, and using only the previous skip connection and the optimization approach of this paper cannot fully utilize the features learned by the encoder. To solve this problem, we consider adding classification labels to guide the encoder to learn more refined features, while using a small number of segmentation labels to achieve accurate segmentation of medical images; (2) Although our method has fewer number parameters and GFLOPs, the FPS does not achieve a significant advantage over other methods, and in the future, we will consider improving the efficiency of the model without loss of accuracy; (3) Although we have demonstrated the effectiveness of the proposed method through quantitative and qualitative results, more specific work on model interpretability is lacking in the field of medical image segmentation. In the future, we will work on methods with better interpretability and more reasonable performance; (4) SEAformer has demonstrated outstanding performance in medical image segmentation, laying a solid foundation for its application in other fields, such as salient object detection and binary image segmentation. In the future, we will consider further extending SEAformer to other domains and evaluating its performance to explore its adaptability and effectiveness in different applications.

6. Conclusion

In this work, we propose SEAformer, a dual-encoder network with a selective edge aggregation Transformer, aimed at addressing challenges of large-scale variation in lesions and blurred boundaries in medical images. Unlike traditional CNN-based encoders, our Transformer encoder efficiently captures the global contextual information necessary for medical image segmentation. In contrast to other dual-encoder approaches, SEAformer enables feature interaction at each stage, making the learned features complementary. The proposed SEA module allows the network to focus on the target boundary without adding extra learning parameters. Furthermore, existing segmentation networks often underutilize the features learned by the encoder. To address this, we introduce a combined boundary loss function and adopt a multi-level optimization strategy. By incorporating upsampling modules and loss functions across multiple scales of the encoder, we achieve multi-scale feature fusion, enhancing the encoder's ability to capture features at different resolutions. This optimization strategy also promotes the network's learning of boundary-related features during encoding, thereby accelerating convergence and improving segmentation performance. Our qualitative and quantitative analyses of segmentation results on five publicly available medical datasets demonstrate SEAformer's effectiveness in localizing targets with significant scale variations and complex boundaries. Additionally, the proposed framework is general and flexible, performing well on smaller datasets and showing potential for broader applications in medical image segmentation tasks.

CRediT authorship contribution statement

Jingwen Li: Writing – original draft, Project administration, Conceptualization. **Jilong Chen:** Investigation, Formal analysis, Data curation. **Lei jiang:** Visualization, Validation, Software. **Ruoyu Li:** Writing – original draft, Software, Resources. **Peilun Han:** Writing – review & editing, Methodology, Investigation. **Junlong Cheng:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgment

This work was supported by the Chengdu Key R&D Program: 2022-YF05-02120-SN.

Appendix. Supplementary material

Our algorithm is designed to retain feature maps with significant boundary information as the output of the CNN branch, and feature maps that do not meet expectations are discarded. The algorithm is implemented by simply traversing each channel and comparing it with a threshold value.

Algorithm 1 Channel Selection Algorithm

Input: Average aggregated X_{EEB} channel orientation feature maps: $\tilde{m} \in R^{H\times W\times 1}$, and the result of element-wise multiplication of X_{SFS} and X_{EEB}: $M \in R^{H\times W\times C}$

Output: Feature map after channel selection: X_{out}

Initialize: T is an empty list
for $c \leftarrow 0$ to $C-1$ **do**
  $m \leftarrow M[..., c]$
  if $\sum_{j=0}^{W-1}\sum_{i=0}^{H-1} m_{(i,j)} > \sum_{j=0}^{W-1}\sum_{i=0}^{H-1} \tilde{m}_{(i,j)}$ **then**
    $T[c] \leftarrow 1$
  else
    $T[c] \leftarrow 0$
  end if
end for
return $T \odot M$

Data availability

Data will be made available on request.

References

[1] Kelei He, Chen Gan, Zhuoyuan Li, Islem Rekik, Zihao Yin, Wen Ji, Yang Gao, Qian Wang, Junfeng Zhang, Transformers in medical image analysis: A review, Intell. Med. (2022).

[2] Evan Shelhamer, Jonathon Long, Trevor Darrell, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2016) 1–1.

[3] Olaf Ronneberger, Philipp Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 9351, 2015, pp. 234–241.

[4] Junlong Cheng, Shengwei Tian, Long Yu, Shijia Liu, Chaoqing Wang, Yuan Ren, Hongchun Lu, Min Zhu, DDU-net: A dual dense U-structure network for medical image segmentation, Appl. Soft Comput. 126 (2022) 109297.

[5] Simon Jégou, Michal Drozdzal, David Vázquez, Adriana Romero, Yoshua Bengio, The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation, in: Proceedings of the IEEE CVPR Workshops, 2017, pp. 1175–1183.

[6] Md. Zahangir Alom, Chris Yakopcic, Mahmudul Hasan, Tarek Taha, Vijayan Asari, Recurrent residual U-net for medical image segmentation, J. Med. Imaging 6 (2019).

[7] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, Jianming Liang, Unet++: A nested U-net architecture for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 11045, 2018, pp. 3–11.

[8] Tiange Xiang, Chaoyi Zhang, Dongnan Liu, Yang Song, Heng Huang, Weidong Cai, Bio-net: Learning recurrent bi-directional connections for encoder-decoder architecture, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 12261, 2020, pp. 74–84.

[9] Yu Liu, Yu Shi, Fuhao Mu, Juan Cheng, Xun Chen, Glioma segmentation-oriented multi-modal MR image fusion with adversarial learning, IEEE/CAA J. Autom. Sin. 9 (8) (2022) 1528–1531.

[10] Junlong Cheng, Shengwei Tian, Long Yu, Hongchun Lu, Xiaoyi Lv, Fully convolutional attention network for biomedical image segmentation, Artif. Intell. Med. 107 (2020) 101899.

[11] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, Yuyin Zhou, TransUNet: Transformers make strong encoders for medical image segmentation, 2021, CoRR abs/2102.04306.

[12] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, Vishal M. Patel, Medical transformer: Gated axial-attention for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 12901, 2021, pp. 36–46.

[13] Zhiqin Zhu, Mengwei Sun, Guanqiu Qi, Yuanyuan Li, Xinbo Gao, Yu Liu, Sparse dynamic volume TransUNet with multi-level edge fusion for brain tumor segmentation, Comput. Biol. Med. (2024) 108284.

[14] Yang Xu, Kun Yu, Guanqiu Qi, Yifei Gong, Xiaolong Qu, Li Yin, Pan Yang, Brain tumour segmentation framework with deep nuanced reasoning and swin-T, IET Image Process. 18 (6) (2024) 1550–1564.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, Alan L. Yuille, DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 834–848.

[16] Zhiqin Zhu, Xianyu He, Guanqiu Qi, Yuanyuan Li, Baisen Cong, Yu Liu, Brain tumor segmentation based on the fusion of deep semantics and edge information in multimodal MRI, Inf. Fusion 91 (2023) 376–387.

[17] Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, Yong Man Ro, Structure boundary preserving segmentation for medical image with ambiguous boundary, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2020, pp. 4816–4825.

[18] Jiacheng Wang, Lan Wei, Liansheng Wang, Qichao Zhou, Lei Zhu, Jing Qin, Boundary-aware transformers for skin lesion segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 12901, 2021, pp. 206–216.

[19] Kun Wang, Xiaohong Zhang, Xiangbo Zhang, Yuting Lu, Sheng Huang, Dan Yang, EANet: Iterative edge attention network for medical image segmentation, Pattern Recognit. 127 (2022) 108636.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, ICLR, 2021.

[21] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, Shuicheng Yan, MetaFormer is actually what you need for vision, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022, pp. 10819–10829.

[22] Huisi Wu, Shihuai Chen, Guilian Chen, Wei Wang, Baiying Lei, Zhenkun Wen, FAT-net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327.

[23] Yuanyuan Li, Ziyu Wang, Li Yin, Zhiqin Zhu, Guanqiu Qi, Yu Liu, X-net: A dual encoding–decoding method in medical image segmentation, Vis. Comput. (2023) 1–11.

[24] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, Bjoern H. Menze, Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields, in: Medical Image Computing and Computer-Assisted Intervention, MICCAI, vol. 9901, 2016, pp. 415–423.

[25] Chen Zhao, Yan Xu, Zhuo He, Jinshan Tang, Yijun Zhang, Jungang Han, Yuxin Shi, Weihua Zhou, Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images, Pattern Recognit. 119 (2021) 108071.

[26] Jinquan Zhang, Zhuang Luan, Lina Ni, Liang Qi, Xu Gong, MSDANet: A multi-scale dilation attention network for medical image segmentation, Biomed. Signal Process. Control 90 (2024) 105889.

[27] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, Jiang Liu, CE-net: Context encoder network for 2D medical image segmentation, IEEE Trans. Med. Imaging 38 (10) (2019) 2281–2292.

[28] Zhiqin Zhu, Ziyu Wang, Guanqiu Qi, Neal Mazur, Pan Yang, Yu Liu, Brain tumor segmentation in MRI with multi-modality spatial information enhancement and boundary shape correction, Pattern Recognit. 153 (2024) 110553.

[29] Hui Tang, Yuanbin Chen, Tao Wang, Yuanbo Zhou, Longxuan Zhao, Qinquan Gao, Min Du, Tao Tan, Xinlin Zhang, Tong Tong, HTC-net: A hybrid CNN-transformer framework for medical image segmentation, Biomed. Signal Process. Control 88 (2024) 105605.

[30] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, H. Fu, H2Former: An efficient hierarchical hybrid transformer for medical image segmentation, IEEE Trans. Med. Imaging 42 (2023) 2763–2775.

[31] Junlong Cheng, Chengrui Gao, Fengjie Wang, Min Zhu, SegNetr: Rethinking the local-global interactions and skip connections in U-shaped networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2023.

[32] Yundong Zhang, Huiye Liu, Qiang Hu, TransFuse: Fusing transformers and CNNs for medical image segmentation, in: Medical Image Computing and Computer Assisted Intervention, MICCAI, vol. 12901, 2021, pp. 14–24.

[33] Xianyu He, Guanqiu Qi, Zhiqin Zhu, Yuanyuan Li, Baisen Cong, Litao Bai, Medical image segmentation method based on multi-feature interaction and fusion over cloud computing, Simul. Model. Pract. Theory 126 (2023) 102769.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: IEEE/CVF International Conference on Computer Vision, ICCV, 2021, pp. 9992–10002.

[35] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, Zhi-Hua Zhou, Selective convolutional descriptor aggregation for fine-grained image retrieval, IEEE Trans. Image Process. 26 (6) (2017) 2868–2881.

[36] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, Hervé Jégou, ResMLP: Feedforward networks for image classification with data-efficient training, 2021, CoRR abs/2105.03404.

[37] Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2017.

[38] Alexey Bokhovkin, Evgeny Burnaev, Boundary loss for remote sensing imagery semantic segmentation, in: Advances in Neural Networks, ISNN, vol. 11555, 2019, pp. 388–401.

[39] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018.

[40] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang, Swin-Unet: Unet-like pure transformer for medical image segmentation, 2021, CoRR abs/2105.05537.

[41] Junlong Cheng, Shengwei Tian, Long Yu, Chengrui Gao, Xiaojing Kang, Xiang Ma, Weidong Wu, Shijia Liu, Hongchun Lu, ResGANet: Residual group attention network for medical image classification and segmentation, Med. Image Anal. 76 (2022) 102313.

[42] Matt Berseth, ISIC 2017 - skin lesion analysis towards melanoma detection, 2017, CoRR abs/1703.00523.

[43] Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, Jorge Rozeira, PH2 - a dermoscopic image database for research and benchmarking, in: International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, 2013, pp. 5437–5440.

[44] Yutong Xie, Jianpeng Zhang, Yong Xia, Chunhua Shen, A mutual bootstrapping model for automated skin lesion segmentation and classification, IEEE Trans. Med. Imaging 39 (7) (2020) 2482–2493.

[45] Korsuk Sirinukunwattana, Josien P.W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Böhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R.J. Snead, Nasir M. Rajpoot, Gland segmentation in colon histology images: The glas challenge contest, Med. Image Anal. 35 (2017) 489–502.