scientific reports

Check for updates

OPEN

Multi-scale conv-attention U-Net for medical image segmentation

Peng Pan[1,2], Chengxue Zhang[1,2], Jingbo Sun[1]✉ & Lina Guo[1,2]

U-Net-based network structures are widely used in medical image segmentation. However, effectively capturing multi-scale features and spatial context information of complex organizational structures remains a challenge. To address this, we propose a novel network structure based on the U-Net backbone. This model integrates the Adaptive Convolution (AC) module, Multi-Scale Learning (MSL) module, and Conv-Attention module to enhance feature expression ability and segmentation performance. The AC module dynamically adjusts the convolutional kernel through an adaptive convolutional layer. This enables the model to extract features of different shapes and scales adaptively, further improving its performance in complex scenarios. The MSL module is designed for multi-scale information fusion. It effectively aggregates fine-grained and high-level semantic features from different resolutions, creating rich multi-scale connections between the encoding and decoding processes. On the other hand, the Conv-Attention module incorporates an efficient attention mechanism into the skip connections. It captures global context information using a low-dimensional proxy for high-dimensional data. This approach reduces computational complexity while maintaining effective spatial and channel information extraction. Experimental validation on the CVC-ClinicDB, MICCAI 2023 Tooth, and ISIC2017 datasets demonstrates that our proposed MSCA-UNet significantly improves segmentation accuracy and model robustness. At the same time, it remains lightweight and outperforms existing segmentation methods.

Keywords Medical image segmentation, Multi-scale learning, Adaptive convolution, Convolutional attention mechanism

Accurate segmentation of medical images is one of the critical technologies to assist doctors in diagnosing and developing personalized treatment plans[1]. In medical image analysis, the segmentation task is crucial. It is used to identify and extract essential regions in the image and provides in-depth image feature information to help doctors make more accurate diagnostic judgments and treatment decisions. Accurate medical image segmentation is a fundamental part of disease detection, diagnosis, and treatment planning and one of the core challenges in medical imaging applications[2,3]. High-quality segmentation results can provide an essential basis for disease assessment, surgical planning, and image-guided surgical operations. However, traditional manual segmentation is labour-intensive and time-consuming, which makes automated segmentation methods particularly important. Fast and accurate automated segmentation can significantly reduce costs and improve the efficiency and accuracy of diagnosis and treatment[4].

In image processing tasks, U-net[5] is widely used as a classical encoder-decoder network architecture. It adopts an encoder-decoder structure, where the encoder generates low-resolution features, and the decoder up-samples the features and gradually recovers the spatial resolution. The jump-join mechanism in U-net helps to recover the spatial information lost in the pooling layer in the decoding stage, ensuring that the spatial details are well-preserved in the U-net architecture. However, the jump connection in U-net is more direct and only partially utilizes the multilayer feature fusion, which has some limitations. To overcome this limitation, UNet++[6] introduces a nested U-net architecture, which realizes more profound feature fusion through multi-level jump connections, aiming to improve segmentation accuracy and fine-grained information retention by comprehensively utilizing multi-scale features. Moreover, the attention mechanism combined with U-net has also been widely noticed, such as Attention U-Net[7], which suppresses the irrelevant regions of the image and highlights the salient features of the local regions through Attention Gate. However, after adding the attention mechanism with the increase of parameters, the computational amount increases significantly; the research began to approach how to accomplish the task efficiently, such as the lightweight network EGE-UNet[8] maintains the segmentation effect while reducing the computational amount to improve the efficiency.

[1]College of Technology and Data, Yantai Nanshan University, Yantai 265713, China. [2]Peng Pan, Chengxue Zhang and Lina Guo contributed equally to this work. ✉email: sunjingbo2022@163.com

In summary, several of the methods above aim to compensate for the loss of spatial information due to the encoder downsampling process in U-shaped networks and improve the focus on localized regions. Although these methods demonstrate exemplary performance in medical image segmentation tasks, they still need to improve in efficiently extracting multi-scale image features globally. Specifically, U-net and UNet++ use a linear connectivity mechanism. At the same time, Attention U-Net introduces a nonlinear attention-gating mechanism, and these methods mitigate the semantic gap between high- and low-level features to some extent. However, there are still limitations in capturing multi-scale spatial information. Based on the in-depth exploration of this theoretical problem, a new need ensues: how to efficiently extract spatial and channel information of high-level and low-level features in the encoder and decoder in order to better bridge the semantic gap between the encoder and the decoder and to compensate for the loss of information due to downsampling? To this end, we revisit the jump-connection mechanism of U-shaped networks and propose a novel deep-learning network that combines the MSL, AC, and Conv-Attention modules[9]. The new network design aims to enhance further the feature extraction capability of jump connections to capture multi-scale and multi-morphology spatial and channel information more efficiently, thus providing a more accurate and efficient solution for medical image segmentation tasks. Our contributions are summarized as follows:

- In order to improve the segmentation accuracy of the model in complex structured images, the AC module designed in this paper dynamically adjusts the parameters of the convolution kernel in a data-driven way through the combination of the adaptive convolutional layer, batch normalization layer and ReLU activation layer. The adaptive feature of the AC module enhances the flexibility and adaptability of the features so that it can capture the subtle changes in the image while maintaining the original structural information.
- To balance context awareness and detail capture, the Conv-Attention module reduces the computational cost of the attention mechanism by learning by proxy from the downsampling results of the last layer. It retains the essential contextual information and accurately recognizes critical regions in the image.
- This paper introduces the MSL module to better deal with the variation of organizational structure and size in medical images and improve the accuracy of segmentation. The MSL module adopts a step-by-step up-sampling strategy and is able to capture fine-grained features and high-level semantic information in the encoding stage, thus realizing effective modelling of complex structures.The MPE layer included in the MSL module is used to up-sample the feature maps layer-by-layer. In contrast, the MSV layer further fuses the multi-scale features to realize the joint expression of multi-level information.

The improved U-net model proposed in this paper utilizes the global context capture capability of the Conv-Attention module, the multi-scale feature enhancement capability of the MSL module, and the dynamic adaptive capability of the AC module so that the model exhibits higher robustness and accuracy in medical image segmentation tasks. The experimental results show that compared with the traditional U-net and other segmentation models, the network proposed in this paper provides an effective segmentation scheme for medical image processing with significant improvement in segmentation accuracy, edge detail capture, and a significant reduction in computation.

Related work

CNN for medical image segmentation

U-net adopts an encoder–decoder design based on convolution, where features are extracted by gradually compressing the spatial dimensions of the input image through the encoder and then recovering the spatial information step by step through the decoder to generate a high-resolution output. The core advantage of U-net lies in the efficient extraction of cross-layer features, which is realized by directly transferring the features between the corresponding layers of the encoder and decoder through jumping connections and realizing the integration of information from different layers. The fusion of different levels of information is realized. This design is particularly suitable for medical image segmentation tasks, which can capture subtle information in the image while maintaining high accuracy[10,11]. Due to its powerful segmentation performance, U-net has been widely used as an indispensable tool in the field of medical image processing and has achieved remarkable success in numerous 2D semantic segmentation tasks[12–14].

As research advances, improved methods based on U-net continue to emerge. For example, UNet++ effectively integrates the four layers of features of U-net by introducing a multi-level feature connection mechanism, enabling the network to autonomously learn the weights of features of different depths to adapt to complex tasks more flexibly. In addition, some models extend U-net to process 3D image data, such as CRANet[15], 3D U-Net[16] and V-net[17]. These models provide substantial advantages in processing 3D medical images (e.g., CT and MRI data) by encoding and decoding the body data in 3D space.

The application of U-net is also gradually expanding to combine with other network architectures to achieve more efficient performance. For example, Unext[18] combines U-net and multilayer perceptron (MLP) to reduce the number of parameters and increase the computational speed while maintaining high performance; WRANet[19] utilizes the discrete wavelet transform to replace the standard downsampling module of the CNN in U-net, which efficiently removes the high-frequency noise information and enhances the network's noise immunity; DualNet[20] further improves the image segmentation performance by using two U-net models. These innovative U-net-derived networks perform excellent segmentation tasks on multiple datasets, such as ISIC2017[21] and CVC-ClinicDB[22]. In addition, novel networks combining U-net with other architectures are emerging. For example, TransUNet combines U-net with Transformer[23] to achieve a balance between global feature modelling and local feature capture; SwinUnet adopts a pure Transformer approach based on the inspiration of the U-net structure to enhance the segmentation performance further; and Cascaded U-Net[24] enhances the segmentation performance by using a cascading strategy of multiple U- Net model cascading strategy to enhance the capture of

multi-scale features; and Uctransnet[25] utilizes the channel cross-attention mechanism to achieve effective fusion of multi-scale encoder features. These improved methods achieve significant segmentation results on datasets such as Glas[26] and MoNuSeg[27], demonstrating their strong adaptability and performance advantages in different scenarios.

Transformer for medical image segmentation

In deep learning image segmentation and recognition tasks, the attention mechanism has been widely used as an effective means to enhance the performance of models[28–30]. The core idea of the attention mechanism is to give the model the ability to focus on key regions so that the model can automatically ignore irrelevant parts when processing complex data, thus enhancing the efficiency and accuracy of feature extraction. Introducing the attention mechanism into the model structure can effectively improve the recognition of detailed regions in the image and optimize the feature expression, especially in medical image processing that requires high-precision segmentation, which plays an important role.

With the application of the attention mechanism, many new models based on attention have been proposed one after another. For example, Attention U-Net introduces the Attention Gate module into the traditional U-net architecture, which enables the network to focus on the target region more effectively, thus significantly improving the segmentation accuracy and robustness. In addition, Vision Transformer (ViT)[31] introduces the Transformer architecture to image classification tasks by applying the global self-attention mechanism to the whole image. It achieves excellent classification results on large datasets such as ImageNet.ViT's global feature capture capability has enabled it to perform highly in natural image processing, but its application is also subject to computational constraints. Computational resources and training data size also limit performance in natural image processing, but its application.

In the field of medical images, to cope with the problems of data scarcity and heterogeneity, Valanarasu et al. proposed the gated axial attention model MedT[32]. This model introduces a specific gating mechanism that better captures feature information in both vertical and horizontal axes to optimize the segmentation accuracy of medical images further. SeTformer[33] significantly improves the computational efficiency while enhancing the performance of the model by replacing the traditional dot-product self-attention (DPSA) with self-optimizing transmission (SeT). At the same time, MASFormer[34] is an innovative variant based on Transformer, which introduces a hybrid attention-spanning mechanism to handle both long-range and short-range dependencies more efficiently, thus further optimizing the feature extraction capability of the model. However, while these approaches enhance the feature extraction capability of the model by adding an attention-spanning mechanism, they do not directly focus on optimizing the U-net structure and tend to increase the complexity and computational cost of the network. In particular, little attention is paid to the feature maps at different scales, often resulting in poor results and bringing structural redundancy and computational complexity.

Methods

Overall architecture of MSCA-UNet

The overall architecture of the MSCA-UNet proposed in this paper is shown in Fig. 1, which adopts a U-shaped encoder–decoder architecture, where we use adaptive convolution instead of the traditional convolution of the U-net model, which allows for better integration of the local contextual information while reducing the model parameters to make the training more efficient. The encoder has four stages, each with our proposed AC module. Specifically, the AC module takes the input $X \in R^{H \times W \times 3}$ through an adaptive convolutional layer for feature extraction and then through a batch normalization layer for final ReLu activation, which, like the traditional U-net network, uses max-pooling to downsize the feature map, again reducing the number of parameters while

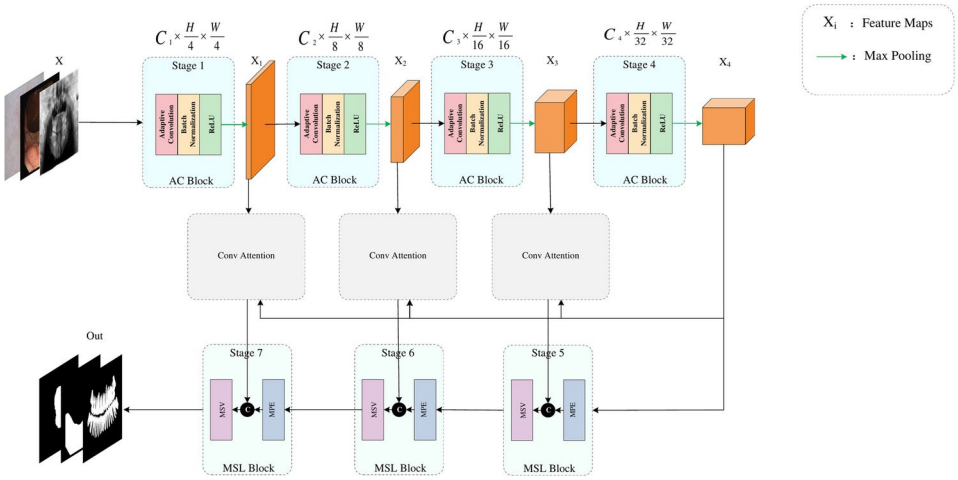

Fig. 1. Our proposed MSCA-UNet is divided into three parts overall. The upper part has four AC modules, and the rectangles between the AC modules represent the feature maps; the middle part corresponds to the Conv-Attention module; the lower part is the MSL module.

extracting significant features. The input enters the decoder after four stages of feature extraction by the encoder. The decoder is divided into three stages, and unlike the U-net network, we propose an MSL module to upsample the feature map and capture the fine-grained multi-scale information and high-level semantic information from the encoding stage. The MPE layer upsamples the features, and the MSV layer further fuses the combined features. For better extraction of meaningful features, we propose the convolutional attention mechanism to weigh the feature channels, which have fewer parameters than the self-attention mechanism.

Adaptive convolution block

In the inference process of ordinary convolution, the parameters of the convolution kernel remain fixed, which limits the adaptive ability to handle diverse input features and makes it difficult to effectively process organs or lesion regions of different morphologies and scales. So this paper proposes a module based on adaptive convolution called AC Module (Adaptive Convolution Module). The module consists of an adaptive convolution layer, a batch normalization layer, and a ReLU activation layer, which can dynamically adjust the parameters of the convolution kernel according to different input features, thus enhancing the network's ability to express features. As shown in Fig. 1, the input features first pass through the adaptive convolutional layer for efficient feature extraction, followed by the batch normalization and activation operations. The process can be described by the following equation.

$$X_A = AdaConv(X), \tag{1}$$

$$X_B = BatchNorm(X_A), \tag{2}$$

$$X_i = ReLU(X_B), \tag{3}$$

where AdaConv stands for adaptive convolution operation, BatchNorm stands for batch normalization, ReLU is the activation function.

In traditional convolutional operations, the kernel parameters remain fixed throughout the inference process, limiting the model's adaptive ability to handle diverse input features. In the AC module, however, each input X prompts the adaptive convolutional layer to adjust dynamically, enabling the model to flexibly capture feature patterns at different scales, angles and viewpoints. The AC module significantly improves the model's feature adaptive ability and extraction efficiency through the synergy of adaptive convolution, batch normalization, and ReLU activation. Meanwhile, optimizing the number of parameters also improves the model's convergence speed and overall stability.

Convolution–attention

To efficiently capture feature maps' global context information, we propose a convolution-like low-dimensional proxy high-dimensional attention mechanism module, the Conv-Attention module,which is shown in Fig. 2. Compared with the standard self-attention mechanism, the Conv-Attention module significantly reduces the computational complexity while maintaining the global sensory field and is suitable for high-resolution feature map processing. Specifically, after the multilayer AC module processes the input feature X, the dimension is reduced to the original 1/32, and this low-dimensional representation serves as a proxy for the high-dimensional features. The Conv-Attention module first performs a 1×1 convolution operation on this low-dimensional feature channel map to obtain the feature X'. It adjusts the number of its channels to be the same as that of the encoder's output feature map at the ith layer. Next, a mapping operation is performed on X' and X_i using weight matrices W_q, W_k, andW_v such that X' serves as Query, while the downsampled representation of X_i serves as Key and Value:

$$Q' = X'W_q, \tag{4}$$

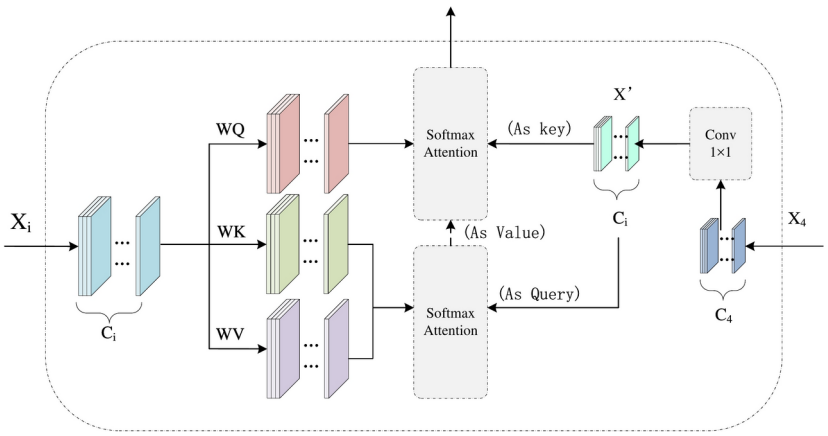

Fig. 2. Details of our proposed Conv-Attention module: different scale feature maps X_i, last layer downsampled X_4, Conv as a convolution operation.

$$K = X_i W_k, \tag{5}$$

$$V = X_i W_v. \tag{6}$$

Then, the Softmax attention mechanism is applied to compute the attention matrix of X' as an agent of X_i:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d^k}}). \tag{7}$$

This matrix is then computed with the Query and Key of X_i by attention to arrive at the final feature matrix with global contextual information:

$$Q = X_1 W_q, \tag{8}$$

$$K' = X' W_k, \tag{9}$$

$$Out = Attention(Q, K', Attention(Q, K, V)). \tag{10}$$

The Conv-Attention module introduces a downsampling strategy to reduce the spatial resolution of the input feature maps before generating the Query, Key, or Value matrices. This strategy downsamples the feature maps to $1/s$ of the original resolution, where s is the downsampling rate. Thus, the attention computation is simplified to the original $1/s^2$ and resource consumption is significantly reduced. Compared with the traditional self-attention module, the Conv-Attention module focuses more on retaining high-level abstract features while capturing long-range dependencies. Assigning weights to the high-level features after multiple rounds of feature extraction with i-layer feature channels makes the model focus more on the discriminative feature channels. On the one hand, this optimizes the expression of multi-scale features and the use of computational resources and improves the recognition ability of the model; on the other hand, it effectively reduces the number of parameters and significantly speeds up the computation. Compared with the standard self-attention mechanism, the Conv-Attention module shows stronger robustness and adaptability in complex scenes.

Multi-scale learning block

The MSL module proposed in this paper is shown in Fig. 3 , aims to up-sample the feature map, combining fine-grained local features and high-level semantic information to capture the multi-scale information in the encoding stage fully. Through multi-scale feature fusion, the MSL module can effectively enhance the model's ability to understand complex scenes, especially in extracting and integrating information at different scales and layers. The MSL module consists of two core layers: the MPE and MSV layers. In the encoder–decoder architecture, the encoding phase gradually compresses the feature maps to extract high-level semantic information, but some fine-grained spatial details may need to be recovered. The primary role of the MPE layer is to up-sample the feature maps to recover the spatial resolution and then use the combination of feature maps with different resolutions to fuse the high-level semantic information with the low-level detailed features. Through multiple up-sampling operations, the MPE layer can generate a set of multi-scale feature maps containing rich spatial details, which are further extracted and fused by depth-separable convolution. At the same time, by recovering layer by layer, the MPE layer can maintain the integrity of the detailed information and combine it with the multi-scale information from the encoding stage to realize fine-grained high-resolution characterization. The MSV layer further performs multi-scale feature fusion based on the up-sampled feature maps in the MPE layer. SS2D[35] provides a new approach for this layer to enhance the characterization capability of features by effectively integrating features from different scales to provide a deeper understanding of the details and semantic information in multi-scale and complex scenes. The MSV layer employs multi-scale convolutional operations to

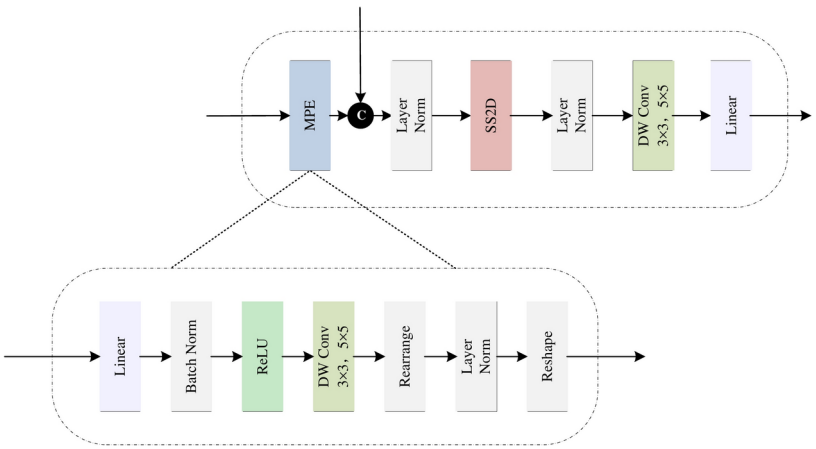

Fig. 3. We propose the MSL module, SS2D for multi-scale extraction of VMamba.

process the up-sampled features to adequately fuse the feature information from different scales and generate a feature map with a high resolution and high quality. The MSV layer uses multi-scale convolutional operations to process the up-sampled features to fully integrate the feature information from different scales and generate joint features with stronger robustness. By combining the MPE and MSV layers, the MSL module efficiently captures and reconstructs multi-scale features from the encoding stage and integrates fine-grained details and high-level semantic information. This feature is particularly suitable for task scenarios such as medical image segmentation that require precise localization and recognition. Compared with traditional up-sampling methods, the MSL module focuses more on combining multi-scale features, enabling the model to cope with scale variations and detail loss in complex scenes. Through multi-scale learning and feature fusion, the MSL module effectively improves the feature map's spatial resolution and characterization ability and enhances the model's ability to express fine-grained and global semantics.

Hybrid loss function

The commonly used loss functions in medical image segmentation tasks mainly include classification loss and segmentation loss[36]. To improve the accuracy of medical image segmentation, this paper uses a combination of cross-entropy loss and Dice loss as the overall loss function. Cross-entropy loss quantifies the difference between the predicted values and the true values, while dice loss evaluates the consistency between the predicted results and the true annotations.

$$L_{CE} = -\sum_{i} T_i \times \log(P_i), \tag{11}$$

$$L_{DC} = 1 - \frac{2 \times \sum_i P_i \times T_i}{\sum_i T_i} + \sum_i P_i, \tag{12}$$

where T_i is the true label using ONE-HOT coding and P_i is the predicted probability. Although Dice loss is well adapted to class-imbalanced data, it is still challenging to achieve excellent segmentation performance in network training using Dice loss alone. In order to segment the tumour region more accurately, this paper combines the advantages of the two loss functions, which can effectively balance the classification performance and segmentation accuracy to optimize the performance of the model, defined as follows:

$$Loss = \alpha L_{CE} + \beta L_{DC}, \tag{13}$$

where α and β are hyperparameters in the hybrid loss function, in this paper $\alpha = 0.4$ $\beta = 0.6$ is selected,this weight configuration enhances the model's ability to learn small - target boundary features by reducing the background weight ($\beta = 0.6$). Meanwhile, the 0.4:0.6 weight ratio, while maintaining the symmetry of the Dice coefficient, dynamically adjusts the loss contributions of the foreground and background, further strengthening the model's learning of foreground details. Additionally, this strategy suppresses background noise interference, improves the model's adaptability to cross-modal data distribution differences, and ensures the segmentation stability of the model under different imaging conditions. Therefore, it provides an interpretable parameter optimization solution for automated segmentation tasks in complex clinical scenarios.

Experiments and results

In this section, we first introduce the three public datasets. The details of the experiments with evaluation metrics, while later we validate the segmentation performance of MSCA-UNet on ISIC2017[21], MICCAI 2023 Tooth[37], and CVC-ClinicDB[22] by comparing it with U-net[5], Attention U-Net[7], DualNet[20], MedT[27],TransUNet[22] Unext[12], EGE-UNet[8], WRANet[13] SeTformer[32] are evaluated. Finally, ablation experiments on our model are outlined.

Datasets

CVC-ClinicDB

This publicly available dataset contains 612 colonoscopy images from real colonoscopy videos, with a resolution of 384x288 pixels.

MICCAI 2023 tooth

This dataset is a high-resolution medical image containing 2D scanned X-ray images acquired from a dental imaging device, which is an accurate pixel-level segmentation of the main structures of the teeth, such as crowns and roots[37]. The training set is 2000 images, and the test set is 500.

ISIC2017

This dataset contains 2000 dermoscopic images with detailed pixel-level segmentation annotations[21]. The training, validation, and test sets are 1279, 150, and 600 images, respectively. Each image has a high resolution and is suitable for nuanced feature analysis.

To standardize the input style across different datasets, all images were resized to a uniform size of 256 × 256 before training. Data enhancement techniques, including random rotation, flipping, scaling, and cropping, were also employed to increase the diversity of the data and enhance the robustness of the model.

In the following two subsections, we elaborate on the details of the experiment and the evaluation metrics and conclude with a discussion of the results.

Experimental details

In this work, we apply the proposed model to the CVC-ClinicDB, MICCAI 2023 Tooth and ISIC2017 datasets and compare it with the standard U-net and SOTA models. The specific experimental setup is as follows:

Model training

All experiments are conducted using the same computational resources and environment. The Adam optimizer is used, and the initial learning rate is set to 0.001, and the learning rate is gradually reduced using the cosine annealing strategy to accelerate convergence. The number of training rounds was set to 100, and the Early Stopping strategy was used (Early Stopping) to avoid overfitting. Each dataset is divided into a training set and a test set. The model is trained on the training set, and the optimal model is selected and evaluated on the test set to ensure the reliability of the experimental results and the model's generalization ability.

Experimental platforms

All experiments were conducted on a computer equipped with an NVIDIA RTX 4090 GPU, and the deep learning framework was PyTorch.

Evaluation metrics

In this study, to comprehensively evaluate the segmentation performance of the model, we use the Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) as the primary evaluation metrics. These two metrics are mainly used to measure the degree of overlap between the predicted segmentation results and the accurate segmentation results, which are defined as follows.

DSC

The Dice coefficient is used to measure the overlap between the predicted segmentation region and the real region and is defined as:

$$DSC = \frac{2 \times |P \cap T|}{|P| + |T|}, \tag{14}$$

where P denotes the segmentation region predicted by the model, and T denotes the actual segmentation region. The DSC is between 0 and 1, with higher values indicating a better match between the segmentation results and the accurate segmentation.

IoU

IoU is used to evaluate the proportion of the intersection of the predicted region with the actual region in the concatenation set, which is defined as:

$$DSC = \frac{|P \cap T|}{|P \cup T|}. \tag{15}$$

Similarly, IoU values range from 0 to 1, with higher values indicating a better match between the segmented region predicted by the model and the actual segmented region.

FLOPs

The computational power of the model is usually measured by the number of floating point operations (FLOPs, Floating Point OMPErations), which indicates the number of floating point operations that the model needs to perform during the inference process. The higher the FLOPs, the higher the computational demand of the model.

Results and discussion

To comprehensively validate the effectiveness of our proposed method, we conducted experiments on three representative public datasets: the CVC-ClinicDB, the MICCAI 2023 Tooth, and the ISIC2017. Seven state-of-the-art methods are employed to validate the performance of MSCA-UNet. The experimental results show that our method performs well in coping with image segmentation tasks of different complexity levels, demonstrating the proposed module's significant advantages in enhancing feature extraction and segmentation performance.

Comparison on CVC-ClinicDB dataset

In order to verify the effectiveness of our proposed module in dealing with boundary fuzzy problems, we choose CVC-ClinicDB intestinal polyp detection for testing. The eight models were tested thrice, taking the mean plus or minus the variance to represent the final results. The experimental results are shown in Table 1. Our method outperforms the other models on the CVC-ClinicDB dataset, with DSC and IoU scores achieving 89.77% and 82.87% on average. Mainly, it leads MedT by nearly 22% and 27% in DSC and IoU, respectively, which is a significant improvement.

Comparison on MICCAI 2023 tooth dataset

Faced with a structurally complex image situation, we use MICCAI 2023 Tooth teeth for testing. The experimental results are shown in Table 2. Our method leads in both evaluation metrics, achieving an average of 92.37% and 88.24% in DSC and IoU scores, reflecting its superior performance for complex image processing.

Method	DSC	IoU
U-net[5]	88.42 ± 0.32	81.22 ± 0.54
Attention U-Net[7]	88.44 ± 0.04	81.32 ± 0.18
DualNet[19]	85.74 ± 0.15	77.19 ± 0.10
MedT[26]	67.56 ± 0.03	55.76 ± 0.18
TransUNet[22]	89.07 ± 0.03	81.75 ± 0.18
Unext[12]	79.30 ± 0.08	69.39 ± 0.19
EGE-UNet[8]	80.86 ± 0.45	71.39 ± 0.77
WRANet[13]	89.14 ± 0.08	82.08 ± 0.09
SeTformer[32]	89.43 ± 0.10	82.31 ± 0.05
Ours	89.77 ± 0.24	82.87 ± 0.13

Table 1. Comparisons with state-of-the-art models on CVC-ClinicDB dataset.

Method	DSC	IoU
U-net[5]	91.71 ±0.01	87.14 ± 0.07
Attention U-Net[7]	91.70 ± 0.01	87.12 ± 0.02
DualNet[19]	90.88 ± 0.01	85.68 ± 0.01
MedT[26]	89.25 ± 0.01	82.30 ± 0.01
TransUNet[22]	89.07 ± 0.03	81.75 ± 0.18
Unext[12]	89.57 ± 0.01	83.11 ± 0.01
EGE-UNet[8]	90.64 ± 0.01	82.99 ± 0.02
WRANet[13]	91.63 ± 0.02	86.93 ± 0.14
SeTformer[32]	91.43 ± 0.10	86.31 ± 0.05
Ours	92.37 ± 0.01	88.24 ± 0.01

Table 2. Comparisons with state-of-the-art models on MICCAI 2023 tooth dataset.

Method	DSC	IoU
U-net[5]	86.67 ± 0.02	78.53 ± 0.02
Attention U-Net[7]	87.45 ± 0.02	79.36 ± 0.05
Dualnet[19]	88.65 ± 0.05	81.61 ± 0.03
MedT[26]	85.83 ± 0.01	77.81 ± 0.01
TransUNet[22]	87.07 ± 0.05	79.85 ± 0.08
Unext[12]	87.52 ± 0.01	80.16 ± 0.05
EGE-UNet[8]	88.05 ± 0.02	80.99 ± 0.01
WRANet[13]	88.50 ± 0.01	81.46 ± 0.01
SeTformer[32]	87.12 ± 0.13	80.56 ± 0.12
Ours	89.42 ± 0.03	82.27 ± 0.03

Table 3. Comparisons with state-of-the-art models on ISIC2017 dataset.

Comparison on the ISIC2017 dataset

We tested comparisons on the ISIC2017 dermatology for situations that are diverse and have more blurred boundaries. The experimental results are shown in Table 3, where our method's DSC and IoU scores achieved 89.42% and 82.27% on average, proving the robustness of image segmentation in the face of situation-complex boundary blurring.

The experimental results benefit from the synergistic effect of our proposed modules: The AC module enhances the ability to capture boundary detail by dynamically adjusting the convolution kernel. The Conv-Attention module effectively directs the model to focus on the key regions and suppresses the background noise. The MSL module integrates the multi-scale information to improve the model's comprehension of the global structure and local details. Together, these modules form an efficient feature extraction framework, which enables the model to exhibit excellent segmentation performance in different complex scenarios.

Visualization of segmentation results

Figures 4, 5 and 6 show the segmentation visualization results for the CVC-ClinicDB, MICCAI 2023 Tooth and ISIC2017 datasets. Our method exhibits higher accuracy in the segmentation task. The first column is the

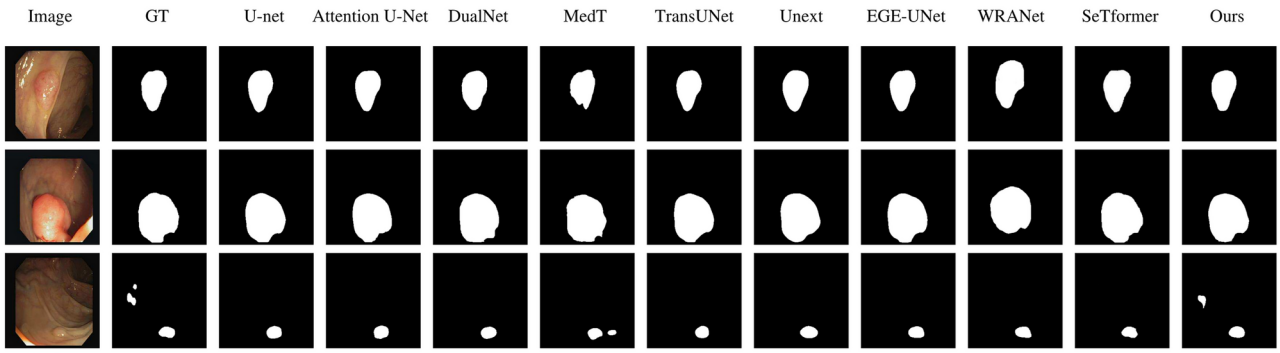

Fig. 4. Comparison chart of experimental results at CVC-ClinicDB.

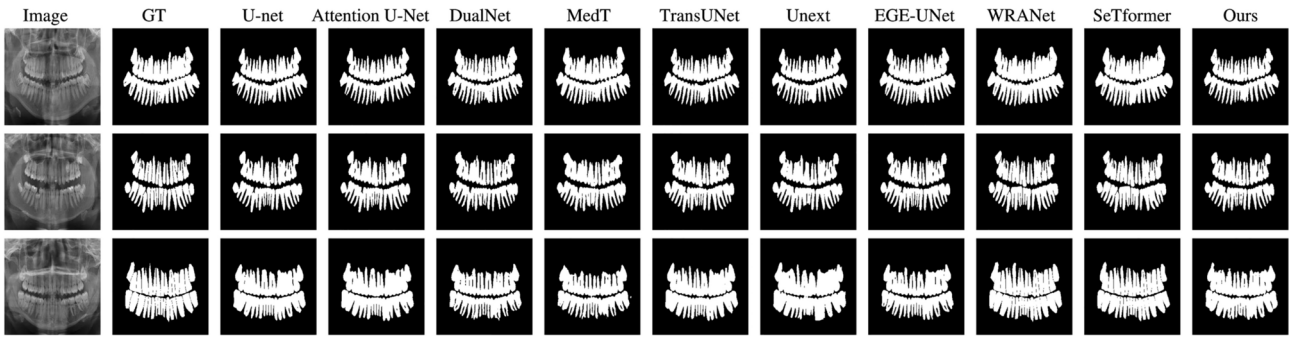

Fig. 5. Comparison chart of experimental results at MICCAI 2023 Tooth.

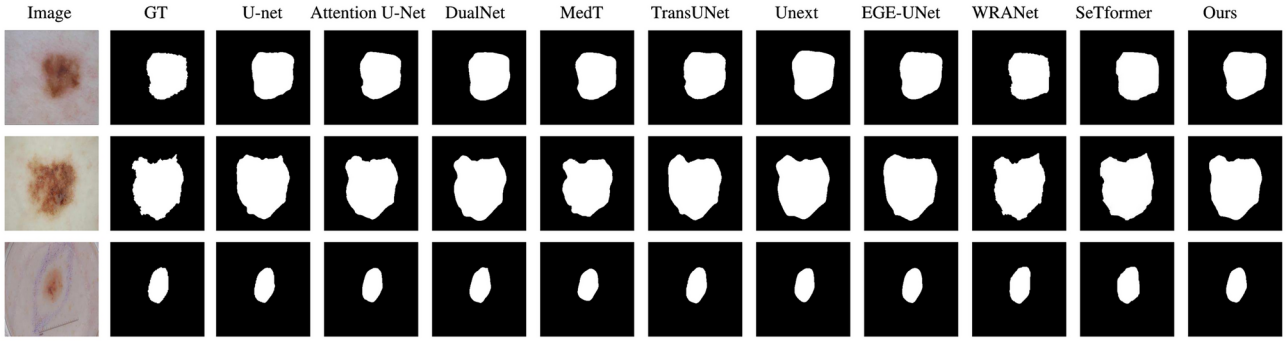

Fig. 6. Comparison chart of experimental results at ISIC2017.

original image, the second column is Ground Truth, and the rest of the columns are U-net, Attention U-Net, DualNet, MedT, TransUNet, Unext, EGE-UNet, WRANet, SeTformer and our method in that order. Thanks to the modules' superiority, our method can accurately segment regions with blurred boundaries while recognizing the target locations of complex structures. For example, in the CVC-ClinicDB dataset, our method can capture the edges of intestinal polyps more clearly, whereas the other models suffer from more missegmented regions. In the MICCAI 2023 Tooth dataset, our method performs well in processing complex structural images, accurately distinguishing the boundaries between teeth.

Moreover, in the ISIC2017 dataset, facing dermatologic images with diverse situations and blurred boundaries, our method can capture more detailed features and generate more accurate segmentation results. In contrast, models such as MedT and Unext are slightly less capable of capturing details. The results show that our module is efficient, and our method can effectively learn detailed features in the data, which is a significant advantage over other SOTA models.

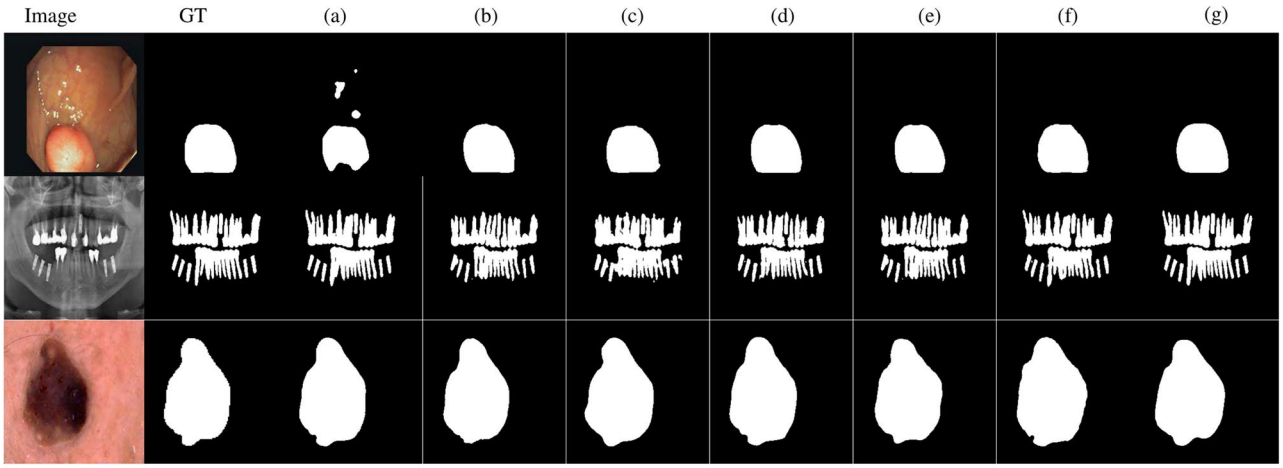

Fig. 7. Visualization of ablation experiments on each of the three datasets: (**a**) +AC, (**b**) +Conv-Attention, (**c**) +MSL, (**d**) +AC+Conv-Attention, (**e**) +AC+MSL, (**f**) +Conv-Attention+MSL, (**g**) +AC+Conv-Attention+MSL.

Method	DSC	IoU
Baseline	87.90	80.34
+AC	88.12	80.44
+Conv-Attention	88.70	80.67
+MSL	87.93	80.41
+AC+Conv-Attention	88.73	81.20
+AC+MSL	88.24	81.17
+Conv-Attention+MSL	88.35	81.42
+AC+Conv-Attention+MSL	**89.83**	**82.72**

Table 5. Ablation experiments on CVC-ClinicDB dataset. Significant values are in bold.

Method	DSC	IoU
Baseline	91.60	86.76
+AC	91.71	86.84
+Conv-Attention	91.97	87.02
+MSL	91.73	86.91
+AC+Conv-Attention	92.30	88.21
+AC+MSL	92.01	87.32
+Conv-Attention+MSL	92.24	87.97
+AC+Conv-Attention+MSL	**92.40**	**88.30**

Table 6. Ablation experiments on MICCAI 2023 tooth dataset. Significant values are in bold.

Ablation experiments

Our method is designed with three key modules: the AC module, the Conv-Attention module, and the MSL module. Comprehensive ablation experiments are performed on multiple datasets to validate and visualize the ablation experiment segmentation results. The experimental results are shown in Tables 5, 6 and 7, and the visualization results are shown in Fig. 7. "Baseline + AC + Conv-Attention + MSL" exhibits optimal performance on all the test datasets, which fully proves the validity of the combination of this module.

First, when the AC, Conv-Attention, or MSL modules are introduced alone, the model's segmentation accuracy on the CVC-ClinicDB and ISIC2017 datasets is significantly improved, and the effect is especially prominent when dealing with boundary-blurring regions. On the MICCAI 2023 Tooth dataset, the model also shows excellent performance in segmenting complex structured regions. In addition, the modules demonstrate unique advantages in mitigating information loss, enhancing contextual modelling capabilities, and refining boundary features.

Upon further testing the combined effect of the modules, we find that the modelling performance is improved by two-by-two combinations, such as "AC + Conv-Attention", "Conv-Attention + MSL", or "AC + MSL",

Method	DSC	IoU
baseline	86.83	78.74
+AC	87.30	79.72
+Conv-Attention	88.34	80.15
+MSL	87.54	79.95
+AC+Conv-Attention	89.31	81.12
+AC+MSL	88.20	80.71
+Conv-Attention+MSL	89.15	82.07
+AC+Conv-Attention+MSL	**89.58**	**82.46**

Table 7. Ablation experiments on ISIC2017 dataset. Significant values are in bold.

Method	FLOPs (M)	Params (M)	Inference time (ms)
U-net[5]	300	31.04	287
Attention U-Net[7]	1100	33.90	312
DualNet[19]	2500	39.9	502
MedT[26]	1710	20.23	732
TransUNet[22]	1590	25.31	577
Unext[12]	10	1.77	**25**
EGE-UNet[8]	**2**	**0.05**	351
WRANet[13]	1100	34.4	305
SeTformer[32]	2100	22.71	610
Ours	850	7.32	343

Table 4. Comparisons with state-of-the-art models on complexity analysis. Significant values are in bold.

which can further improve the segmentation results of the three datasets. Finally, the model achieves the best segmentation performance when all three modules, AC, Conv-Attention and MSL, are used simultaneously. Our method performs excellently in all segmentation tasks on the CVC-ClinicDB and MICCAI 2023 Tooth datasets.

The experimental results show that our proposed modular design can effectively improve the overall performance of the segmentation tasks. This further emphasizes the importance of enhancing feature extraction capabilities and multi-scale information modelling within the encoder-decoder framework, which provides an efficient and robust solution to the complex image segmentation problem.

Complexity analysis

We analyzed the performance of our proposed method with existing SOTA models in terms of the number of parameters, FLOPs, and inference time, and the results are shown in Table 4. Although our method introduces various modules to enhance segmentation performance, it performs better in segmenting intestinal polyps, teeth and skin diseases with a significant reduction in the number of parameters and inference time compared to models such as U-net, MedT and WRANet.Although the UNext model outperforms our model in inference time, our model still achieved superior segmentation performance while ensuring the inference time remains as short as possible, demonstrating superior overall performance. In addition, thanks to the introduction of the efficient AC module and Conv-Attention module, our method can achieve fast inference and excellent segmentation performance while keeping the computational complexity low. This makes it a more attractive choice for practical applications.

Conclusion

This study proposes an improved model based on the U-net backbone network that combines the AC module, the MSL module, and the Conv-Attention mechanism to cope with boundary ambiguity, complex structure, and missing information in medical image processing. Through experimental validation on several public datasets (e.g., CVC-ClinicDB, MICCAI 2023 Tooth, and ISIC2017), our approach achieves significant improvements in accuracy and robustness, significantly outperforming the seven network models tested in both Dice Similarity Coefficient (DSC) and Intersection-to-Union Ratio (IoU) metrics.The experimental results show that our model exhibits significantly better segmentation performance than comparable methods on multi-modal video datasets, X-ray datasets in the field of medical imaging, and MRI datasets, verifying its generalization ability across modalities and domains. Notably, the model maintains stable performance in low-contrast X-ray images and high-noise video frames, further demonstrating its robustness and generalization ability, and providing a reliable solution for automated segmentation tasks in complex clinical scenarios.

Compared with the traditional U-net, our model reduces the number of parameters and computational complexity and improves the segmentation accuracy without degrading the performance. The introduced AC and MSL modules effectively enhance the feature extraction capability, and the Conv-Attention mechanism

further efficiently improves the model's focus on critical features. Overall, the method is suitable for accurate segmentation of medical images and provides new ideas and feasible solutions for image processing tasks in other complex scenes.

In future work, we plan to further optimize the computational efficiency of the model and explore its applicability in more medical domains. We expect to validate its generalization performance and practical application value on a wider range of datasets.

Data availability

All datasets are publicly available. CVC-ClinicDB https://www.kaggle.com/datasets/balraj98/cvcclinicdb, MICCAI 2023 Tooth https://tianchi.aliyun.com/dataset/156596, and ISIC2017 https://challenge.isic-archive.com/data/#2017.

Received: 27 November 2024; Accepted: 26 March 2025
Published online: 08 April 2025

References

1. Siddique, N., Paheding, S., Elkin, C. P. & Devabhaktuni, V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access* **9**, 82031–82057 (2021).
2. Liu, X., Song, L., Liu, S. & Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability* **13**(3), 1224 (2021).
3. Ramesh, K., Kumar, G. K., Swapna, K., Datta, D. & Rajest, S. S. A review of medical image segmentation algorithms. *EAI Endorsed Trans. Pervas. Health Technol.* **7**(27), 6 (2021).
4. Azad, R. et al. Medical image segmentation review: The success of u-net. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 1 (2024).
5. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-assisted intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18* 234–241 (Springer, 2015).
6. Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N. & Liang, J. UNet++: A nested U-net architecture for medical image segmentation. https://arxiv.org/abs/1807.10165 (2018).
7. Oktay, O. et al. Attention U-net: Learning where to look for the pancreas. https://arxiv.org/abs/1804.03999 (2018).
8. Ruan, J., Xie, M., Gao, J., Liu, T. & Fu, Y. EGE-UNet: An efficient group enhanced UNet for skin lesion segmentation. https://arxiv.org/abs/2307.08473 (2023).
9. Muksimova, S., Umirzakova, S., Shoraimov, K., Baltayev, J. & Cho, Y.-I. Novelty classification model use in reinforcement learning for cervical cancer. *Cancers* **16**(22), 3782 (2024).
10. AnbuDevi, M. & Suganthi, K. Review of semantic segmentation of medical images using modified architectures of unet. *Diagnostics* **12**(12), 3064 (2022).
11. Li, X. et al. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* **37**(12), 2663–2674 (2018).
12. Xu, G., Zhang, X., He, X. & Wu, X. Levit-unet: Make faster encoders with transformer for medical image segmentation. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)* 42–53 (Springer, 2023).
13. Guan, S., Khan, A. A., Sikdar, S. & Chitnis, P. V. Fully dense unet for 2-d sparse photoacoustic tomography artifact removal. *IEEE J. Biomed. Health Inform.* **24**(2), 568–576 (2019).
14. Weng, Y., Zhou, T., Li, Y. & Qiu, X. Nas-unet: Neural architecture search for medical image segmentation. *IEEE Access* **7**, 44247–44257 (2019).
15. Zhao, Y., Wang, S., Ren, Y. & Zhang, Y. Cranet: A comprehensive residual attention network for intracranial aneurysm image classification. *BMC Bioinform.* **23**(1), 322 (2022).
16. Çiçek, Ö. et al. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17–21, 2016, Proceedings, Part II 19* 424–432 (Springer, 2016).
17. Milletari, F., Navab, N. & Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* 565–571 (IEEE, 2016).
18. Valanarasu, J. M. J. & Patel, V. M.: Unext: Mlp-based rapid medical image segmentation network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 23–33 (Springer, 2022).
19. Zhao, Y., Wang, S., Zhang, Y., Qiao, S. & Zhang, M. Wranet: Wavelet integrated residual attention u-net network for medical image segmentation. *Complex Intell. Syst.* **9**(6), 6971–6983 (2023).
20. Pham, Q., Liu, C. & Hoi, S. *DualNet: Continual Learning, Fast and Slow*. https://arxiv.org/abs/2110.00175 (2021).
21. Codella, N. C. F. et al. *Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)*. https://arxiv.org/abs/1710.05006 (2018).
22. Bernal, J. et al. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs saliency maps from physicians. *Comput. Med. Imaging Graph.* **43**, 99–111 (2015).
23. Vaswani, A. et al. *Attention Is All You Need*. https://arxiv.org/abs/1706.03762 (2023).
24. Jiang, Z., Ding, C., Liu, M. & Tao, D. Two-stage cascaded u-net: 1st place solution to brats challenge 2019 segmentation task. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I 5* 231–241 (Springer, 2020).
25. Wang, H., Cao, P., Wang, J. & Zaiane, O. R.: Uctransnet: Rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2441–2449 (2022)
26. Sirinukunwattana, K., Snead, D. R. & Rajpoot, N. M. A stochastic polygons model for glandular structures in colon histology images. *IEEE Trans. Med. Imaging* **34**(11), 2366–2378 (2015).
27. Kumar, N. et al. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Trans. Med. Imaging* **36**(7), 1550–1560 (2017).
28. Han, K. et al. Transformer in transformer. *Adv. Neural. Inf. Process. Syst.* **34**, 15908–15919 (2021).
29. Han, K. et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(1), 87–110 (2022).
30. Parmar, N. et al. Image transformer. In *International Conference on Machine Learning* 4055–4064 (PMLR, 2018).
31. Dosovitskiy, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (ICLR, 2021)
32. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I. & Patel, V. M. *Medical Transformer: Gated Axial-Attention for Medical Image Segmentation*. https://arxiv.org/abs/2102.10662 (2021).
33. Shamsolmoali, P., Zareapoor, M., Granger, E. & Felsberg, M. *SeTformer is What You Need for Vision and Language*. https://arxiv.org/abs/2401.03540 (2024).

34. Zhang, Q., Ram, D., Hawkins, C., Zha, S. & Zhao, T. *Efficient Long-Range Transformers: You Need to Attend More, but Not Necessarily at Every Layer*. https://arxiv.org/abs/2310.12442 (2023).
35. Liu, Y. et al. *VMamba: Visual State Space Model*. https://arxiv.org/abs/2401.10166 (2024).
36. Jadon, S. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* 1–7 (IEEE, 2020).
37. Wang, Y. et al. STS MICCAI 2023 challenge: Grand challenge on 2d and 3d semi-supervised tooth segmentation. *CoRR*. https://doi.org/10.48550/ARXIV.2407.13246 (2024).

Author contributions

Jingbo Sun handles project management and communication, Peng Pan focuses on theory and algorithm development, Chengxue Sun works on literature review and model training, while Lina Guo executes experiments, visualizes data, and interprets results.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-96101-8.

Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.