



OPEN Dual-branch dynamic hierarchical U-Net with multi-layer space fusion attention for medical image segmentation

Zhen Wang¹, Shuang Fu^{2✉}, Hongguang Zhang², Chunyang Wang³, Chunhui Xia², Pen Hou², Chunxue Shun² & Ge Shun²

Accurate segmentation of organs or lesions from medical images is essential for accurate disease diagnosis and organ morphometrics. Previously, most researchers mainly added feature extraction modules and simply aggregated the semantic features to U-Net network to improve the segmentation accuracy of medical images. However, these improved U-Net networks ignore the semantic differences of different organs in medical images and lack the fusion of high-level semantic features and low-level semantic features, which will lead to blurred or miss boundaries between similar organs and diseased areas. To solve this problem, we propose Dual-branch dynamic hierarchical U-Net with multi-layer space fusion attention (D2HU-Net). Firstly, we propose a multi-layer spatial attention fusion module, which makes the shallow decoding path provide predictive graph supplement to the deep decoding path. Under the guidance of higher semantic features, useful context features are selected from lower semantic features to obtain deeper useful spatial information, which makes up for the semantic differences between organs in different medical images. Secondly, we propose a dynamic multi-scale layered module that enhances the multi-scale representation of the network at a finer granularity level and selectively refines single-scale features. Finally, the network provides guiding optimization for subsequent decoding based on multi-scale loss functions. The experimental results on four medical data sets show D2HU-Net enables the most advanced segmentation capabilities on different medical image datasets, which can help doctors diagnose and treat diseases

Medical image segmentation provides important help for doctors to diagnose diseases, analyze pathology, design treatment plans and arrange surgical plans¹. However, manual annotation and segmentation of medical image diagnosis face great challenges in clinical practice. Firstly, the morphological diversity of the lesion region, the similarity between the gray level and the adjacent tissue, and the image noise, under the influence of subjective factors, manual segmentation is prone to over-segmentation and under-segmentation. In addition, medical images have small volume, low contrast, unclear adjacent boundaries, and significant differences in lesion areas between different patients or even between the same patient. Finally, different imaging principles of medical imaging devices lead to different clarity of medical images, which will also lead to low accuracy of segmentation². Therefore, there is an urgent need to exploit advanced computer vision techniques to develop a model that can quickly and accurately identify, localize, and segment target regions in medical images.

U-Net is a revolutionary innovation in the field of medical image segmentation³. It is the encoder-decoder architecture. There are many skip connections from low to high levels between the encoder and the decoder. It can well combine low-resolution and high-resolution information to make up for the information lost in down sampling. It allows U-Net to obtain more accurate pixels, which improves the utilization of information and the accuracy of segmentation. Influenced by U-Net, scholars have developed many variants based on U-Net network and tried their best to improve U-Net network⁴. The improvement of these algorithms is mainly reflected in three aspects: convolution operation⁵, skip connection⁶, and depth module⁷.

Convolution operation is the basic building block of convolutional neural networks⁸. Classical convolution operation can make full use of context information to map and extract features from data⁹. scholars introduced the streamlined attention mechanism of variable size and convolution, which fully understood the volume

¹School of Public Health, Qiqihar Medical University, Qiqihar 161003, China. ²College of Pharmacy, Qiqihar Medical University, Qiqihar 161003, China. ³Qiqihar First Hospital, Qiqihar 161003, China. ✉email: fsjt1980@qmu.edu.cn

context, and flexibly distorted the sampling grid, so that the model could appropriately adapt to different data patterns.

Skip connection is a breakthrough in U-Net network, which aggregates low-level features in encoder and high-level features in decoder. Previous skip connections simply aggregate low-order and high-order features by concatenation between encoders or decoders¹⁰. TransAttUnet uses additional multi-scale skip connections between decoder blocks to aggregate upsampled features with different semantic scales. In this way, the ability to represent multi-scale context information can be enhanced and discriminative features can be generated¹¹.

The last aspect is the depth block, which is generally the backbone of the network, which also determines the feature inference mode of the network. It has been proved that the performance of the network gradually increases with the increase of depth and width¹². However, as the depth increases, the difficulty of training also increases. To alleviate the shortcomings of these depth modules, several improved networks have been developed using residual blocks¹³, dense blocks¹⁴, and initial residual blocks¹⁵. In order to expand the global feature extraction ability of the network, the attention mechanism is introduced into the depth block. Such as attention gate¹⁶, dual attention¹⁷, spatial and channel attention¹⁸. In addition, Transformer, from the field of natural language processing, has excellent performance by using self-attention mechanism to obtain data features¹⁹. Quite a few researchers have created many medical graph segmentations using Transformer structure or convolutional network structure²⁰. They have shown acceptable results.

Although the above scholars have achieved good results in the field of medical image segmentation, there are still some shortcomings in the segmentation process. Firstly, the common convolutional structure in U-Net network can only extract features from a single aspect, and the semantic features of images are discontinuous. Secondly, the aggregation strategy of skip join treats the context information equally, lacks the combination of high-level and low-level features, and cannot make full use of context dependence. This also leads to unclear segmentation of the boundary of the target region. Finally, most methods are designed for specialized lesions and organs with a single image modality. Since the features of multimodal images are very different from each other, the segmentation results are not ideal for different lesions and organs and different modalities.

During the design of the model, two key challenges were encountered. First, extracting semantic features from the target region at multiple scales proved difficult, particularly when the medical image contrast was low, the target region shape was irregular, distribution imbalanced, and semantic information fragmented. Second, effectively fusing low-level and high-level semantic information while leveraging their complementary nature was necessary. To address the first challenge, we propose a dynamic multi-scale hierarchical module that enhances the network's multi-scale representation capability at finer granularities. This module selectively refines single-scale features to improve noise resistance. To address the second challenge, we introduce a multilayer spatial attention fusion module, which integrates feature information across both channel and spatial dimensions. It also establishes multidimensional interactions between these dimensions. The module dynamically aggregates various types of semantic features, allowing shallow decoding prediction maps to complement deeper decoding paths. This enables D2HU-Net to select useful contextual features from low-level information and extract high-level semantic and detailed information at different scales under the guidance of high-level features. Moreover, it bridges the semantic differences between organs across different medical images. Simultaneously, the multi-scale features are selectively refined to extract more relevant spatial information. By capturing contextual information of the target, attention features across different scales are effectively fused, addressing boundary blurring and missing regions in target segmentation.

The main contributions of this paper can be summarized as follows:

1. A dual-branch dynamic hierarchical U-Net (D2HU-Net) with multi-layer spatial fusion attention is proposed, in which the shallow decoding path provides guidance for the deep decoding path and improves the generalization ability.
2. A multi-layer spatial attention fusion module is proposed. It not only focuses on the distinguishing characteristics of channel and spatial dimension, but also establishes the interdimensional interaction between channel and spatial dimension. In addition, the weighted parameters are adapted to further blend the features of each view. Finally, selective nuclear feature aggregation can dynamically aggregate advanced features of different branches/types, effectively improving the accuracy of medical image segmentation.
3. Dynamic multi-scale layered module is proposed, selectively utilize multi-scale features integrated at different scales, and refine features at each scale. Enhance the ability of multi-scale expression of network features. Improve the precision of segmentation of the target area.
4. We experiment D2HU-Net on four multi-modal medical image datasets and obtain high segmentation accuracy. The proposed model has good robustness.

Related works

CNN-based segmentation networks

In the early medical image segmentation, feature engineering and deep learning²¹ were used to extract image features. With the development of deep learning algorithms, in the absence of any manual features, deep learning-based methods can automatically extract various features from medical images in an end-to-end form for medical image segmentation. Long et al. designed the end-to-end convolutional network architecture FCN for semantic segmentation for the first time²². Influenced by FCN, U-Net architecture is proposed for medical image segmentation. U-Net adopts a strategy of skipping connections, integrating low-level features with high-level features, thereby preserving some detailed features. In the later work, scholars have proposed various improved semantic segmentation models based on U-Net network. Xiao et al. proposed Res-U-Net by introducing the weighted attention mechanism and residual connections into U-Net²³. Around the same time, Radiuk et al. proposed that a U-shaped network with nested and dense jump connections was proposed for multi-organ

segmentation²⁴. However, these methods only collect features from neighborhood pixels and lack awareness of global image information. In addition, ICNet is proposed as a cascading network architecture to connect multi-resolution feature maps with appropriate label guidance and model compression²⁵. Chen et al. designed GLNet by combining two CNN branches to preserve both the global background and the detailed local information²⁶. These branches interact with each other through a depth feature mapping sharing strategy, resulting in better performance. However, since convolution kernels tend to extract only local regions, this approach often has limitations when modeling remote relationships.

To solve this problem, several methods have been proposed where extended convolution is used instead of step convolution to expand the acceptance field and model non-local information²⁷. In addition, some scholars have proposed to introduce attention mechanism into CNN to improve the network's focus on features. For example, Wang et al. applied self-attention modules in deep neural networks to explore the validity of non-local dependencies in images, sequences, and videos²⁸. CBAM²⁹ and DANet³⁰ enhance the representation ability of CNN by using the attention mechanism along channel dimension and spatial dimension respectively. These methods have achieved remarkable performance on different tasks. Despite the commendable achievements of the aforementioned scholars, the prevailing convolutional structure is incapable of accommodating multiple perspectives, and the semantic features extracted from the graph exhibit discontinuity. Consequently, this paper proposes a dynamic multi-scale hierarchical module to enhance the multi-scale representation of the network at a finer granularity level, thereby enhancing the segmentation accuracy of the network.

Despite the commendable achievements of the aforementioned scholars, the prevailing convolutional structure is incapable of accommodating multiple perspectives, and the semantic features extracted from the graph exhibit discontinuity. Consequently, this paper proposes a dynamic multi-scale hierarchical module to enhance the multi-scale representation of the network at a finer granularity level, thereby enhancing the segmentation accuracy of the network.

Attention mechanisms

Attention mechanisms were originally designed for natural language models. Since attention mechanism can effectively solve the problem that convolution operation cannot highlight the object features and suppress the noise in the network, attention mechanism has become a research hotspot and has been widely applied in the field of computer vision. Wang et al. introduced non-local operations into the spatial attention mechanism to obtain spatial dependencies by calculating the attention force³¹. This is a spatial attention mechanism that focuses on where the effective features are in the feature map. In addition, there are channel attention mechanisms, which are mainly concerned with what are the effective features in the feature graph and suppress the redundant channels of the feature graph. Hu et al. proposed a channel attention mechanism to learn global context features by recalibrating channel dependencies on different channels by squeezing and exciting blocks³². At the same time, Woo et al. proposed the convolutional block attention module (CBAM) to coordinate the joint action of channels and spatial attention mechanisms³³. In recent years, several multi-scale attention mechanisms have emerged. Chen et al. proposed Transattunet to design multi-level guided attention, integrating self-conscious attention modules from self-attention with transform and global Spatial attention into TransAttunet to effectively learn non-local interactions between encoder features³⁴. Huang et al. proposed a cross-concern module in which the network can adaptively capture remote context information from image dependencies³⁵.

However, these methods based on spatial attention mechanisms and channel attention mechanisms are independent dependencies. Therefore, the interaction between the spatial dimension and the channel dimension in the network is limited. multi-layer spatial attention fusion module proposed in this paper can capture the interactive features of each dimension, focus on the important features of the feature graph from four dimensions, and improve the segmentation ability of the network.

Methods

In this section, firstly, D2HU-Net will be briefly introduced. Secondly, the dynamic multi-scale layered module is introduced in detail, which is the basic model of D2HU-Net. Thirdly, the attention mechanism of multi-layer space fusion is introduced theoretically. Finally, the calculation process of the loss function is introduced in detail.

Overview of the proposed network

The main purpose of medical image segmentation is to segment the target region from various formats of medical images. In order to realize this process, D2HU-Net based neural network is proposed for medical image segmentation. The specific schematic diagram of D2HU-Net is shown in Fig. 1. Specifically, D2HU-Net contains one encoding path and two decoding paths, thus forming a two-branch image segmentation architecture.

The main function of the coding stage is to encode the input medical images into semantic features. The encoding phase consists of a total of five coding blocks, referred to as E_1, E_2, E_3, E_4, E_5 (from top to bottom). The encoder is composed of dynamic multi-scale layered modules, which are described in detail in “Multi-layer spatial attention fusion module”. Behind each encoder block, a downsampling operation using maximum pooling is applied, which serves to double the number of channels for the semantic features.

At the end of the encoder E_4 encoding, the shallow decoding path begins to perform semantic segmentation from bottom to top. Among them, the shallow decoding path includes three decoders, which are D_3^1, D_1^2, D_1^1 from bottom to top. Both the decoder structure and the encoder structure are composed of dynamic multi-scale layered modules.

Compared to the maximum pooling of the encoder, the decoder is always present, and an upsampling operation by a transposed convolution of 4×4 increases the space size of the semantic feature by a factor of 2, while doubling the channel size of the feature. The input of the decoder is connected to the output of the encoder

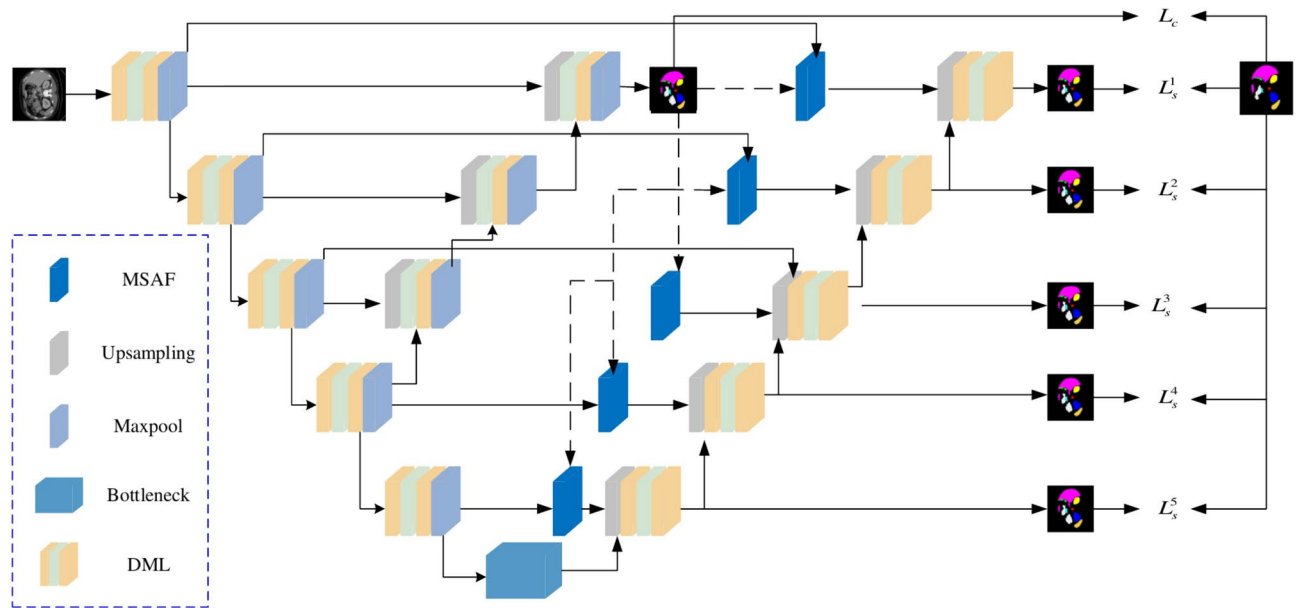


Fig. 1. Architecture of the proposed D2MU-Net for medical image segmentation.

at the same scale with the sampled features on the previous decoder. Finally, after the convolution of 1×1 , the segmentation diagram M_s of the shallow path is obtained, which is denoted as. The shallow split path includes 4 encoders and 3 decoders, and the flow is: $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4 \rightarrow D_1^3 \rightarrow D_1^2 \rightarrow D_1^1$.

The segmentation result of the deep decoding path is the final medical image segmentation result graph, which includes 5 decoders $D_2^i, i = 1, 2, \dots, 5$. The decoder of the deep decoding path and the decoder of the shallow decoding path have the same structure. The jump connection of the deep decoding path is not simply directly connected with the output of the decoder at the same scale, but the target feature in the semantic feature is amplified through the multi-layer spatial attention fusion module. The output of the multi-layer spatial attention fusion module is connected with the decoding semantic feature after being upsampled, and then input into the decoder. There is a great difference between shallow semantic features and deep semantic features. Shallow semantic features contain more detailed information, while deep semantic features include more semantic information. The multi-layer spatial attention fusion module fuses useful information from deep semantic features and shallow semantic features, making the network pay more attention to specific information.

Each decoder is classified, and the segmentation graph for each scale is denoted as $M_s^i, i = 1, 2, 3, \dots, 6$. This segmentation graph is used to define the multi-layer loss function, which improves the stability of training and the robustness of the model. Finally, the final output diagram of D2HU-Net is obtained through the encoding stage $E_1 \rightarrow E_2 \rightarrow E_3 \rightarrow E_4 \rightarrow E_5$, bottleneck and deep decoding path $D_2^5 \rightarrow D_2^4 \rightarrow D_2^3 \rightarrow D_2^2 \rightarrow D_2^1$.

The multi-layer spatial attention mechanism and dynamic multi-scale feature extraction structure of D2HU-Net have been shown to increase the number of parameters and computational complexity of the model while improving the segmentation performance. On larger datasets, the training of the model increases accordingly, especially when computational resources are limited, which may affect the efficiency of practical applications. However, the experimental results demonstrate that appropriately increasing the model's complexity can effectively enhance the segmentation accuracy, particularly in the context of medical images with multi-scale and complex backgrounds. The model exhibits an improved ability to extract important features.

Dynamic multi-scale layered module

Dynamic multi-scale layered module (DML) is an important basic module of D2HU-Net. DML is composed of two dynamic convolution layers and a layered module. The specific diagram is shown in Fig. 2, Each dynamic convolution is connected to the batch normalization and Rectified Linear Unit (ReLU) activation functions. The schematic diagram of dynamic convolution is shown in Fig. 2. In form, $\text{Conv}_i (i=1, 2, \dots, K)$ is the K convolution core of dynamic convolution, so the mathematical formula for dynamic convolution can be defined as follows:

$$\begin{cases} [\pi_1, \pi_2, \dots, \pi_k] = f_\pi(x) \\ y = \sum_{i=1}^k \pi_i \cdot \text{Conv}_i(x) \end{cases} \quad (1)$$

where, x and y , respectively, dynamically convolve the input and output, f_π indicating that the semantic features of the input pass through the global average pooling (GAP), the fully connected layer (FC), the ReLU activation function, the fully connected layer (FC), and finally the weight π is generated by the Softmax function. In traditional convolution operations, the convolution parameters are static and the same for all inputs, while dynamic convolution adaptively learns the weight value π for each input. In summary, each input is represented

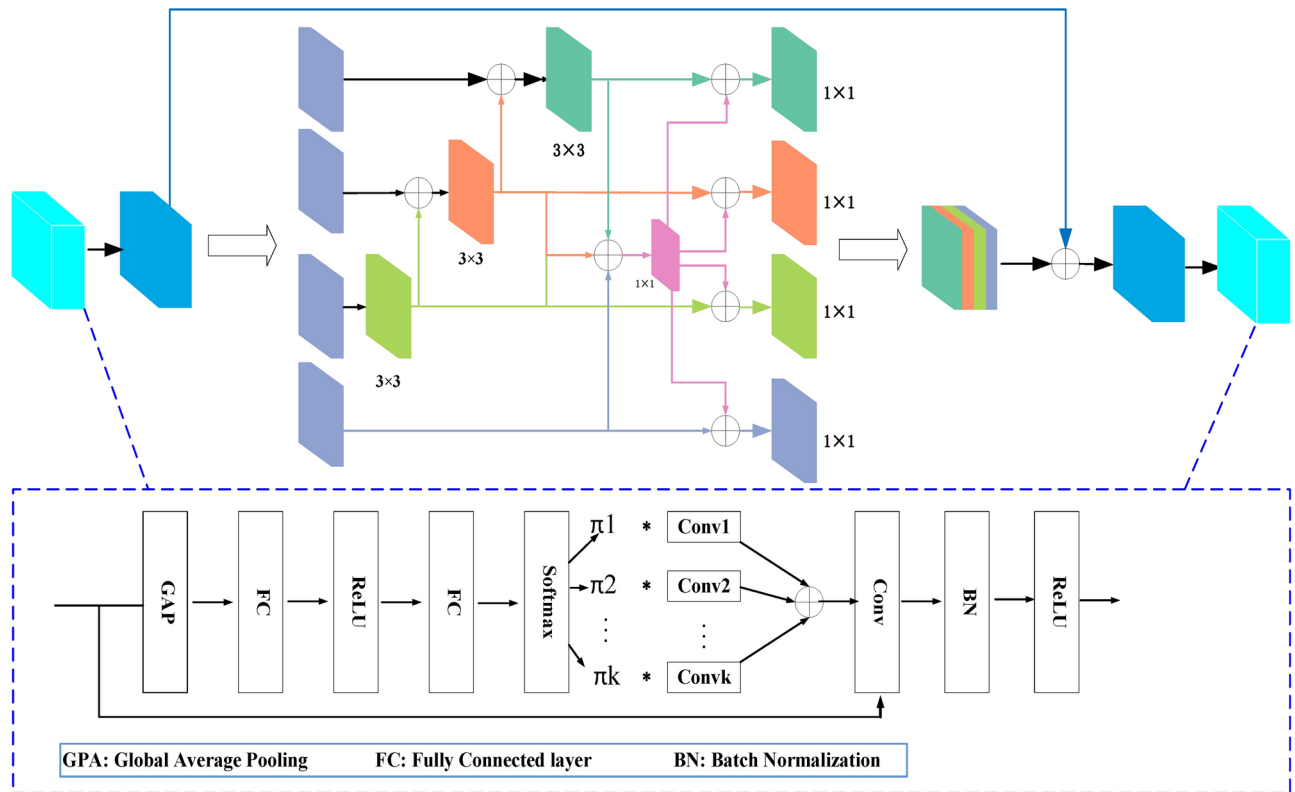


Fig. 2. Architecture of the dynamic multi-scale layered module.

by a unique set of convolution kernels, so the expressiveness of the network is improved, and the parameters of the model are not increased. For parameter settings of dynamic convolution, the K of dynamic convolution is set to 8, and the convolution kernel size is set to 3×3 . The semantic features that have undergone dynamic convolution will be evenly divided into 4 groups along the channel dimension, denoted as f_1, f_2, f_3, f_4 . Except that the number of channels becomes a quarter of the original, the other feature scales are the same as the input. expect f_1 , the rest of f_i are summed element-wise with the output of the 3×3 convolution operation with f_{i-1} , and then batch normalization (BN) and ReLU activation are applied. This process can be described by the following formula:

$$f'_i = \begin{cases} f_i & \text{if } i = 1 \\ \text{Conv}_3(f_i) & \text{if } i = 2 \\ \text{Conv}_3(f_i + f'_{i-1}) & \text{if } i = 3, 4 \end{cases} \quad (2)$$

where Conv_3 represents a convolution layer whose convolution kernel is 3×3 . For these different scales of semantic feature information, we use the method of cross-line fusion information, to exchange and fusion information of (f'_1, f'_2, f'_3, f'_4) , so that each scale of feature information will receive other scales of information. This means that different receiving domains can receive different messages. The specific formula is expressed as follows:

$$f''_i = f'_i + \text{Conv}_1 \left(\sum_{i=1}^4 f'_i \right), \quad i = 1, 2, 3, 4 \quad (3)$$

Since each scale captures multi-scale semantic information in the way of segmentation, in order to solve the problem of weakening information flow caused by segmentation operation, we carry out 1×1 convolution operation for connected semantic features and fuse mixed semantic features of different scales. The formula is as follows:

$$y = \text{Conv}_1 (\text{Concat} (f''_1, f''_2, f''_3, f''_4)) \quad (4)$$

where y represents the output after 1×1 convolution. For the output of features, dynamic convolution operations are again performed, along with batch normalization (BN) and ReLU activation. The maximum pooling operation is then performed as an input to the next encoder.

Multi-layer spatial attention fusion module

Multi-layer spatial attention fusion module (MSAF) is an important part of the decoding of D2HU-Net in the deep decoding path. In order to learn rich context information, multi-scale feature information is aggregated to remove the noise affected by non-target regions. Inspired by³⁶, we propose a space-related module to highlight the features of the target segmentation region step by step. This module is called multi-level spatial attention fusion module.

In contrast to traditional single-attention mechanisms, the proposed multilayer spatial attention fusion module emphasizes spatial features at multiple scales through hierarchical attention mechanisms. This allows the model to capture global context while preserving fine-grained details, improving its accuracy and robustness when processing complex images. Unlike standard channel or spatial attention mechanisms, the multilayer spatial attention fusion module enhances spatial information by performing feature fusion at different levels. This multi-level fusion capability enables the model to analyze images at a finer granularity, effectively balancing both detailed and global information. The specific diagram is shown in Fig. 3.

The input of MSAF is the output result of the encoder at the same scale and the D_1^1 output of the shallow decoding path through channel integration. The multi-level spatial attention fusion module can be divided into two parallel multi-level spatial attention stages and the attention fusion stage. The multi-level spatial attention stage can be divided into four parallel branches, including channel dimension attention, channel and width dimension attention, channel and height dimension attention, and space dimension attention.

Firstly, the first branch takes the output of D_1^1 in the shallow decoding path as an example, expands the number of channels by using 1×1 convolution, and performs down sampling operation at the same time to obtain the same feature size as the output of the encoded block, denoted as $M \in R^{c \times h \times w}$. In the multi-layer spatial attention stage, the first branch focuses on recalibrating channel-level feature representation capabilities. First, we aggregate the spatial features of the input M_c by maximum pooling and average pooling, which can be defined as $X_{1(\max)}^{c \times c} \in R^{c \times 1 \times 1}$ and $X_{1(\text{avg})}^{c \times c} \in R^{c \times 1 \times 1}$. Then, using a multi-layer perceptron (MLP) consisting of two 1×1 convolutional layers and an activation function (Rule), the size of the middle layer is set to $R^{c/2 \times 1 \times 1}$ in order to maintain channel resolution and minimize the number of parameters, and the output of the MLP is summed at the element level. The sum output is then passed through the softmax activation function to get the channel-level attention weight $A_{c \times c}(X) \in R^{c \times 1 \times 1}$. It can be concluded that the mathematical calculation of the weight mapping of the channel-level attention of the first branch is as follows:

$$\begin{aligned} A_{c \times c}(X) &= \theta(\text{MLP}(\text{Maxpool}(X)) + \text{MLP}(\text{Avgpool}(X))) \\ &= \theta(W_2 \xi(W_1(X_{1(\max)}^{c \times c})) + W_2 \xi(W_1(X_{1(\text{avg})}^{c \times c}))) \end{aligned} \quad (5)$$

where θ is the sigmoid function, ξ is the ReLU function, $W_1 \in R^{c/2 \times c}$, and $W_2 \in R^{c \times c/2}$. Finally, the first branch output feature maps $X_1^{c \times c}$ is generated by the following equation:

$$X_1^{c \times c} = A_{c \times c}(X)X \quad (6)$$

The main role of the second branch is to focus on the interaction of channels and height dimensions. First, X is rotated 90 degrees counterclockwise along the height scale to generate a new semantic feature $X_{2r}^{c \times h} \in R^{w \times h \times c}$. Next, feature aggregation of $X_{2r}^{c \times h} \in R^{w \times h \times c}$ is performed using maximum pooling and average pooling, $X_{2r(\max)}^{c \times h} \in R^{1 \times h \times c}$ and $X_{2r(\text{avg})}^{c \times h} \in R^{1 \times h \times c}$, respectively. These outputs are then concatenated with BN using a $K \times K$ convolution operation, and then, using S-type activation function yields the weight mapping of cross attention between channels and height dimensions $A_{c \times h}(X_{2r}^{c \times h}) \in R^{1 \times h \times c}$. In short, its calculation formula is as follows:

$$\begin{aligned} A_{c \times h}(X_{2r}^{c \times h}) &= \theta(f^{k \times k}[\text{Maxpool}(X_{2r}^{c \times h}); \text{Avgpool}(X_{2r}^{c \times h})]) \\ &= \theta(f^{k \times k}[X_{2r(\max)}^{c \times h}; X_{2r(\text{avg})}^{c \times h}]) \end{aligned} \quad (7)$$

where θ is the sigmoid function, $f^{k \times k}$ is the $K \times K$ convolution layer with the BN. The second branch output feature maps $X_2^{c \times h}$ is generated by the following equation:

$$X_2^{c \times h} = \text{Roated}(A_{c \times h}(X_{2r}^{c \times h}))X_{2r}^{c \times h} \quad (8)$$

The third branch is mainly concerned with the interaction between the channel dimension and the width dimension. The process of calculating semantic features is similar to the process of calculating channels and high attention. First, the input X is rotated 90 degrees counterclockwise along the width to obtain a new semantic feature $X_{3r}^{c \times w} \in R^{h \times c \times w}$. After that, perform the same operation as the previous branch. The calculation process of the attention mechanism of channel dimension and width dimension can be summarized as follows:

$$\begin{aligned} A_{c \times w}(X_{3r}^{c \times w}) &= \theta(f^{k \times k}[\text{Maxpool}(X_{3r}^{c \times w}); \text{Avgpool}(X_{3r}^{c \times w})]) \\ &= \theta(f^{k \times k}[X_{3r(\max)}^{c \times w}; X_{3r(\text{avg})}^{c \times w}]) \\ X_3^{c \times w} &= \text{Roated}(A_{c \times w}(X_{3r}^{c \times w}))X_{3r}^{c \times w} \end{aligned} \quad (9)$$

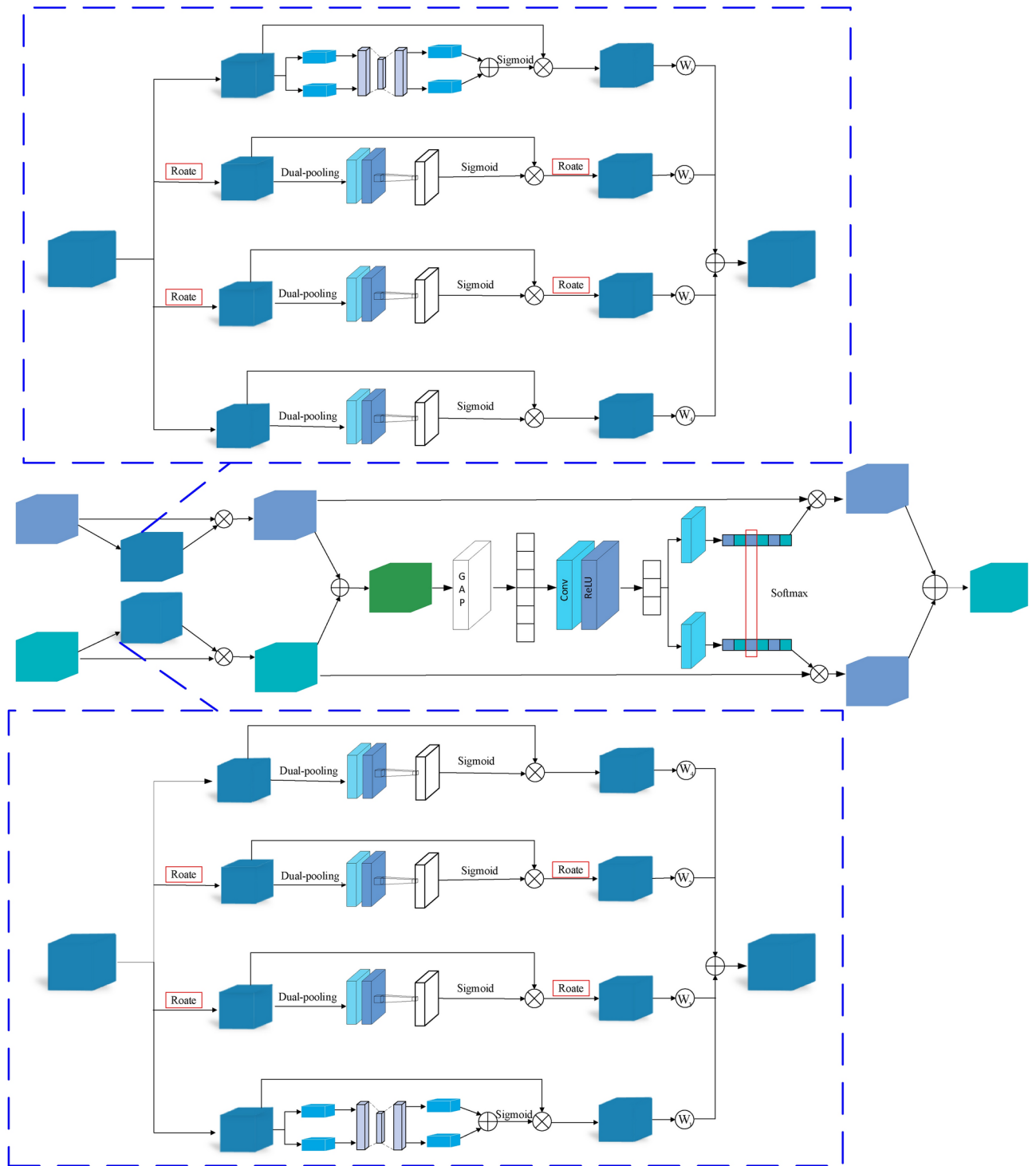


Fig. 3. Architecture of Multi-layer spatial attention fusion module.

The main function of the fourth branch is to focus on the characteristics of the target region in the spatial dimension. The calculation process of the spatial attention map and the output of spatial attention can be summarized as follows:

$$\begin{aligned}
 A_{h \times w} (X_4^{h \times w}) &= \theta \left(f^{k \times k} \left[\text{Maxpool} (X_4^{h \times w}); \text{Avgpool} (X_4^{h \times w}) \right] \right) \\
 &= \theta \left(f^{k \times k} \left[X_{4(\max)}^{h \times w}; X_{3(4\text{avg})}^{h \times w} \right] \right) \\
 X_4^{h \times w} &= A_{h \times w} (X_4^{h \times w}) X_4^{h \times w}
 \end{aligned} \tag{10}$$

Finally, in order to further improve the fusion ability of features, we add a learnable weight parameter to the back of the four branches, so the output in the multi-layer space attention stage can be summarized as:

$$w_i = \frac{\exp(a_i)}{\sum_{j=1}^4 \exp(a_j)}, \quad i = 1, 2, \dots, 4 \quad (11)$$

$$Y = w_1 (X_1^{c \& c}) + w_2 (X_2^{c \& h}) + w_3 (X_3^{c \& w}) + w_4 (X_4^{h \& w}) \quad (12)$$

where, w_i is the normalized weight coefficient and $\sum w_i = 1$; a_i and a_j are the initial weight coefficients.

Similarly, it can be concluded that the output of the multi-layer spatial attention stage of the second branch is Y_2 . After that, we fuse Y and Y_2 through element-level sum operations to obtain the correlation feature graph of high-dimensional semantic features and low-dimensional semantic features, that is $E_{yy} = Y + Y_2$. Then, E_{yy} is transformed into a global statistical descriptor with scale $c \times 1 \times 1$ by GAP operation. Next, we use 1×1 convolutional layers with a ReLU activation function for the channel down sampling operation. To preserve as much channel resolution as possible, the down sampled features $z \in R^{c/2 \times 1 \times 1}$. After that, we convert the feature vector z to $V_Y \in R^{c \times 1 \times 1}$ and $V_{Y_2} \in R^{c \times 1 \times 1}$ using two parallel 1×1 channel upper sampling layers. Finally, we use two weights to re-calibrate the multi-scale feature maps Y and Y_2 dynamically, and accordingly, the final output $M = V_Y Y + V_{Y_2} Y_2$.

loss

In the training process, the loss function consists of using the binary cross-entropy loss L_{BCE} and the dice loss L_{Dice} , where R_{gt} and R_{seg} represent the real label value and the predicted result, respectively. Then L_{BCE} and L_{Dice} can be defined as

$$L_{BCE}(R_{gt}, R_{seg}) = -(1 - R_{gt}) \log(1 - R_{seg}) - R_{gt} \log R_{seg} \quad (13)$$

$$L_{Dice}(R_{gt}, R_{seg}) = 1 - \frac{2R_{gt}R_{seg}}{R_{gt} + R_{seg} + \varepsilon}$$

The ε in the formula is a small constant set to avoid having a zero denominator. Therefore, the loss function in the experiment can be derived according to L_{BCE} and L_{Dice} .

$$L_{seg}(R_{gt}, R_{seg}) = \lambda L_{BCE} + L_{Dice} \quad (14)$$

where λ was set to 0.5 in this experiment.

In order to improve the stability of the training process, the proposed model applies L_{seg} to both decoders in the second stage decoding process to establish a deep supervision loss function. Where M_{GT} represents the true segmentation result graph, so the loss L_s^i of the i -th decoder can be defined as

$$L_s^i(M_{GT}^i, M_s^i) = L_{seg}(M_{GT}^i, M_s^i), \quad i = 1, 2, \dots, 5 \quad (15)$$

In the above formula, M_{GT}^i is defined as

$$M_{GT}^i = \begin{cases} M_{GT} \downarrow_{2^{i-1}} & \text{if } i \geq 2, \\ M_{GT} & \text{if } i = 1. \end{cases} \quad (16)$$

where $\downarrow_{2^{i-1}}$ represents the sub-sample of factor 2^{i-1} . In addition, L_{seg} is also applied to the output of the shallow decoding process, so the loss of the shallow decoding path can be defined as

$$L_c(M_{GT}, M_c) = L_{seg}(M_{GT}, M_c) \quad (17)$$

Finally, the overall loss function during the training of this network can be defined as

$$L_{total} = L_c + \sum_{i=1}^5 L_s^i \quad (18)$$

Experiments

Dataset

Synapse multi-organ segmentation dataset: synapse from the Multi-Graph Abdominal Tagging Challenge held by MICCAI 2015. Different from the traditional single organ segmentation, synapse is the segmentation of eight organs in the abdomen, including the aorta, gallbladder, spleen, left kidney, right kidney, liver, pancreas, spleen, and stomach. The dataset included 30 abdominal clinical CT images, with the number of CT sections contained in each scan ranging from 85 to 192, resulting in a total of 3778 two-dimensional images. The size of all 2D slice images is adjusted to 512×512 , with a voxel spatial resolution of $([0.54 \ 0.54] \times [0.98 \ 0.98] \times [2.5 \ 5.0]) \text{ mm}^3$. In this experiment, 18 samples were divided into training sets, and the remaining 12 were divided into test sets.

Skin Lesion Segmentation: The ISIC 2018 dataset is derived from the ISIC 2018 challenge, which addresses skin lesion segmentation, lesion attribute monitoring, and dermatological classification. The dataset included

2594 images, of which 20.0% were melanoma, 72.0% were moles, and the remaining 8.0% were lipid overflow keratosis. The size of the picture is 512×512 .

Chest X-Ray dataset: This dataset is a pre-chest lung X-ray. This data comes from the National Library of Medicine, National Institutes of Health, Bethesda, MD, USA and Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China³⁷. The dataset consisted of 710 chest X-ray films and accompanying labels. The size of the image is 512×512 .

Kvasir-SEG: The Kvasir-SEG dataset contains 1000 polyp images from the Kvasir dataset v2 and their corresponding ground truths, which were manually seeded by physicians and cross-validated by professional gastroenterologists. The image resolutions included in the Kvasir-SEG range from 332×487 to 1920×1072 pixels, and we adjusted them to 512×512 pixels.

In the process of data preprocessing, we carry out random rotation, horizontal inversion and enlargement of the picture. For each data set, we divided the training set, the validation set, and the test set in a 7:1:2 ratio.

Evaluation metrics

In order to quantitatively analyze the segmentation results, we used five evaluation indicators. including Dice coefficient (Dice), Intersection over Union (IOU), relative volume difference (RVD), average symmetric surface distance (ASSD), maximum symmetric surface distance (MSD). Let R_{gt} and R_{seg} be the ground truth and predicted segmentation result. The mathematical formula for these indicators is as follows:

1. Dice coefficient (DIC): It specifically represents the ratio of the area where two objects intersect to the total area. Its value range is $[0,1]$ and the perfect separation value is 1.

$$DIC = \frac{2(R_{gt} \cap R_{seg})}{R_{gt} + R_{seg}} \quad (19)$$

2. Intersection over Union (IOU): Overlap rate of segmentation result and real result.

$$IOU = \frac{|R_{seg} \cap R_{gt}|}{|R_{seg} \cup R_{gt}|} \quad (20)$$

3. Relative Volume Error (RVD): used to indicate the Volume difference between ground truth and predicted segmentation results. The closer the value is to zero, the higher accuracy of the segmentation accuracy.

$$RAVD = \frac{R_{seg}}{R_{gt}} - 1 \quad (21)$$

4. Average symmetric surface distance (ASSD): the average distance between the surfaces of segmentation results R_{seg} and the ground R_{gt} . Where $d(a, b)$ represents the distance between a and b .

$$ASSD = \frac{1}{|R_{seg}| + |R_{gt}|} \left(\sum_{a \in R_{seg}} \min_{b \in R_{gt}} d(a, b) + \sum_{b \in R_{gt}} \min_{a \in R_{seg}} d(a, b) \right) \quad (22)$$

5. Maximum symmetric surface distance (MSD): The maximum surface distance between the segmentation result R_{seg} and the ground reality R_{gt} , the lower the MSD value, the higher the matching degree between the two sample.

$$MSD = \left(\max_{i \in R_{seg}} \left(\min_{j \in R_{gt}} d(i, j) \right), \max_{i \in R_{gt}} \left(\min_{j \in R_{seg}} d(i, j) \right) \right) \quad (23)$$

Experimental details

All models were based on the Pytorch framework and python 3.8, and all experiments were performed on a deep learning workstation equipped with an Intel Xeon E5-2680 v4 processor, 35 M of L3 cache, 2.4 GHz clock rate, 14 physical cores/28-way multitask processing. In addition, it has 128 GB of DDR4 RAM and an 8 x NVIDIA GeForce RTX 2080Ti super graphics processing unit (GPU) with 11 GB of RAM. Furthermore, the hyperparameters are set to the same for all models, Initial learning rate is 0.0003, Batch size is 16, Epoch is 200, Optimizer is Adam, Growth rate is 0.0001.

To demonstrate the validity of our proposed D2HU-Net, we compare D2HU-Net with nine SOTA segmentation models, including: FCN³⁸, U-Net³, ResUNet³⁹, Attention U-Net⁴⁰, UNet++⁴¹, ResNet++⁴², SWimU-Net⁴³, TransoformU-Net⁴⁴, HiFormer⁴⁵. All comparison methods are set to default parameters. All model weights are retrained on the training set, and the number of training rounds is set to 200. Meanwhile, in order to eliminate the chance of the experiment, we conduct 5 tests on each model on the test set, and select the average value of each evaluation index.

Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
FCN	62.61 ± 0.56	49.67 ± 0.46	− 17.67 ± 0.48	7.98 ± 0.48	43.11 ± 4.2
U-Net	75.73 ± 0.45	64.52 ± 0.52	− 8.07 ± 0.39	5.83 ± 0.39	32.50 ± 3.8
ResU-Net	77.31 ± 0.33	66.40 ± 0.33	− 10.85 ± 0.22	5.49 ± 0.52	31.00 ± 2.5
Attention U-Net	77.57 ± 0.42	66.44 ± 0.28	− 9.91 ± 0.45	5.29 ± 0.37	31.35 ± 2.9
U-Net++	79.33 ± 0.33	67.53 ± 0.36	− 8.68 ± 0.12	4.74 ± 0.28	26.56 ± 5.1
ResU-Net++	76.61 ± 0.24	65.00 ± 0.28	3.21 ± 0.36	4.76 ± 0.42	25.52 ± 2.7
TransformU-Net	79.52 ± 0.14	66.48 ± 0.19	1.63 ± 0.25	4.33 ± 0.37	23.51 ± 3.2
SwimU-Net	76.61 ± 0.31	65.00 ± 0.20	3.21 ± 0.21	4.76 ± 0.22	25.52 ± 2.9
HiFormer	76.92 ± 0.21	65.4 ± 0.36	− 5.56 ± 0.15	4.53 ± 0.38	26.81 ± 2.8
D2HU-Net	80.39 ± 0.19	68.32 ± 0.13	1.54 ± 0.09	4.03 ± 0.36	21.49 ± 2.4

Table 1. Evaluation indicators ± standard deviation of all competing methods in the Synapse multi-organ segmentation dataset.

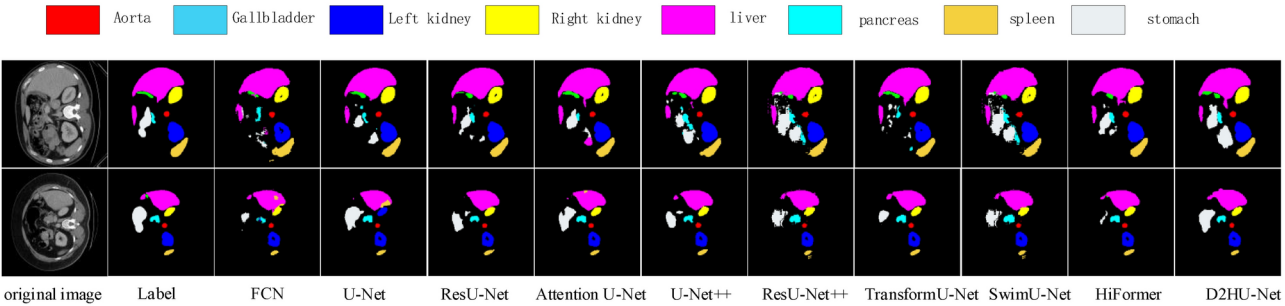


Fig. 4. Visualization of the segmentation results of D2HU-Net and 9 comparison methods on the Synapse multi-organ segmentation dataset.

Experimental results and analysis

In order to demonstrate the superiority of the proposed D2HU-Net model in various aspects, we evaluated the D2HU-Net model quantitatively and qualitatively, comparing it with nine advanced methods on four datasets.

1. Synapse multi-organ segmentation dataset: The quantitative comparison results between the benchmark model of abdominal organ segmentation task in the Synapse multi-organ segmentation dataset and D2HU-Net are shown in Table 1. D2HU-Net is significantly superior to the other 9 CNN-based SOTA methods. In terms of DICE, TransformU-Net achieves the second-best value, while D2HU-Net achieves 0.87% higher than D2HU-Net. For IOU and RAVD, the increase was 0.79% and 0.09%, respectively. For ASSD and MSSD decreased by 0.5 and 2.02, respectively. Specifically, he steadily outperformed most organ segmentation methods. Figure 4 shows the qualitative segmentation results. The results show that this method can segment fine and complex structures accurately, output more accurate segmentation results, and have stronger robustness to complex backgrounds.
2. Skin Lesion Segmentation: The quantitative comparison results between the benchmark model of Skin Lesion Segmentation task and D2HU-Net on the skin lesion segmentation dataset are shown in Table 2. The table of D2HU-Net presented in this paper is ahead of other competitors' methods in 5 evaluation indicators, except ASSD. Specifically, in terms of DICE, D2HU-Net improved by 1.25% and IOU by 0.12% over the most advanced HiFormer, with significant advantages in both RAVD and MSSD. In addition, the variance of D2HU-Net in each evaluation index is also small, which reflects good stability. For qualitative analysis, Fig. 5 shows the visualization of Segmentation results on the Skin Lesion Segmentation dataset. This shows that our proposed D2HU-Net generates more accurate contours than other comparison models in both small and large area targets, and is closest to the label map. In general, D2HU-Net has the best performance, showing good generalization ability and good robustness.
3. Chest X-Ray dataset: The quantitative comparison of the baseline model for Chest X-Ray lung segmentation task in the Chest X-ray dataset with D2HU-Net is shown in Table 3. From the data in the table, we can intuitively see that in addition to RAVD, other indicators of D2HU-Net proposed by us are far ahead of other methods. Due to the simplicity of the data set, the DICE of D2HU-Net is as high as 97.2%, which is 0.3% higher than that of TransformU-Net with the second best value. In terms of IOU, D2HU-Net provided a gain of 0.4%. In addition, there are 0.53 and 3.3 reductions for ASSD and MSSD, respectively. At the same time, the variance is also minimal, indicating that D2HU-Net has good stability. Figure 6 shows some results of qualitative analysis of lung segmentation of X-ray images, indicating that D2HU-Net has achieved

Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
FCN	84.92 ± 0.39	78.61 ± 0.45	13.22 ± 0.45	14.47 ± 4.6	42.77 ± 3.5
U-Net	85.53 ± 0.47	79.83 ± 0.50	14.14 ± 0.36	13.44 ± 3.5	38.77 ± 4.2
ResU-Net	87.87 ± 0.44	80.27 ± 0.69	12.91 ± 0.43	13.27 ± 4.1	37.90 ± 3.6
Attention U-Net	87.68 ± 0.35	79.94 ± 0.43	13.32 ± 0.31	13.52 ± 4.7	38.41 ± 2.7
U-Net++	87.74 ± 0.36	80.06 ± 0.25	13.78 ± 0.28	13.23 ± 3.1	38.37 ± 2.2
ResU-Net++	87.39 ± 0.25	79.51 ± 0.21	14.11 ± 0.31	13.75 ± 2.7	39.08 ± 1.8
TransformU-Net	88.91 ± 0.19	81.75 ± 0.18	- 17.96 ± 0.09	11.21 ± 1.6	50.57 ± 1.7
SwimU-Net	88.26 ± 0.16	80.85 ± 0.21	11.22 ± 0.12	11.80 ± 1.8	31.55 ± 2.1
HiFormer	88.93 ± 0.29	81.89 ± 0.25	15.21 ± 0.13	11.41 ± 2.1	31.74 ± 2.5
D2HU-Net	90.18 ± 0.17	82.01 ± 0.15	5.74 ± 0.21	18.83 ± 1.2	28.96 ± 2.5

Table 2. Evaluation indicators ± standard deviation of all competing methods in the Skin Lesion Segmentation.

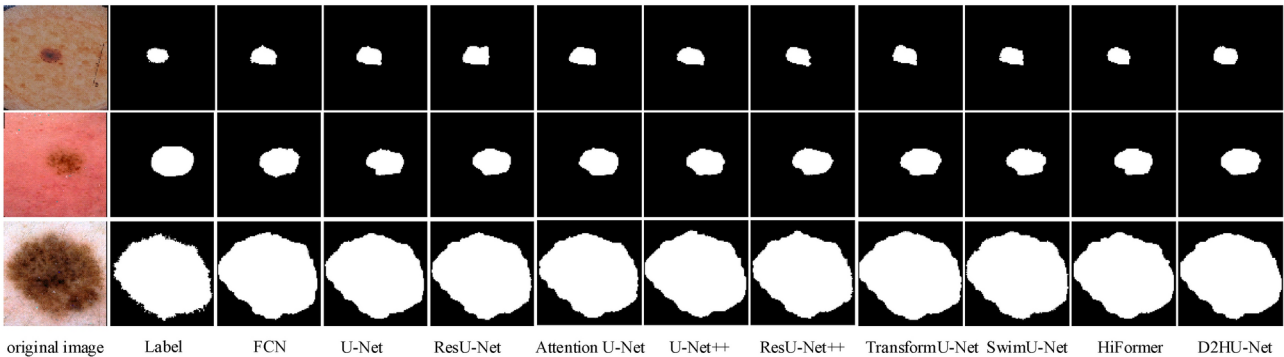


Fig. 5. Visualization of the segmentation results of D2HU-Net and 9 comparison methods on the skin lesion dataset.

Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
FCN	95.6 ± 0.16	93.0 ± 0.29	1.61 ± 0.35	2.75 ± 0.48	11.52 ± 0.44
U-Net	96.4 ± 0.21	93.3 ± 0.22	- 1.93 ± 0.27	2.79 ± 0.44	11.08 ± 0.33
ResU-Net	96.6 ± 0.18	93.5 ± 0.38	- 1.53 ± 0.28	2.67 ± 0.36	10.71 ± 0.36
Attention U-Net	96.4 ± 0.28	93.2 ± 0.25	- 1.96 ± 0.45	2.77 ± 0.45	10.94 ± 0.31
U-Net++	96.6 ± 0.13	93.7 ± 0.21	- 1.41 ± 0.13	2.62 ± 0.27	10.53 ± 0.24
ResU-Net++	95.8 ± 0.22	92.1 ± 0.31	- 2.62 ± 0.21	3.32 ± 0.35	13.23 ± 0.35
TransformU-Net	96.9 ± 0.09	94.1 ± 0.18	- 1.48 ± 0.22	2.36 ± 0.26	9.38 ± 0.22
SwimU-Net	95.5 ± 0.16	91.5 ± 0.20	- 3.55 ± 0.28	3.62 ± 0.22	13.84 ± 0.27
HiFormer	96.6 ± 0.11	93.5 ± 0.17	- 1.1 ± 0.11	2.65 ± 0.25	10.66 ± 0.33
D2HU-Net	97.2 ± 0.08	94.5 ± 0.14	- 3.86 ± 0.17	2.09 ± 0.2	6.08 ± 0.19

Table 3. Evaluation indicators ± standard deviation of all competing methods in the Chest X-Ray dataset.

4. Kvasir-SEG: The quantitative comparison between the baseline model of the rectal polyp segmentation task on the Kvasir-SEG dataset and D2HU-Net is shown in Table 4. Table 4 lists the specific data of each evaluation index, and we can find that all indexes are better than other models. Specifically, D2HU-Net was 0.7% higher in DICE than the second-best TransformU-Net and 1.7% higher in IOU than HiFormer, with ASSD and MSSD decreasing by 0.63 and 0.02, respectively. Figure 7 shows the segmentation results of some polyp images. The main difficulties in the segmentation of gastrointestinal polyps are that the color of polyps and normal tissue is similar, the boundary is not clear, and the illumination is not uniform. Thanks to the Multi-layer spatial attention fusion module, high-level semantic information and low-level semantic information are fused, which also enables D2HU-Net to segment results closer to labels than other methods and accurately extract polyp structure.

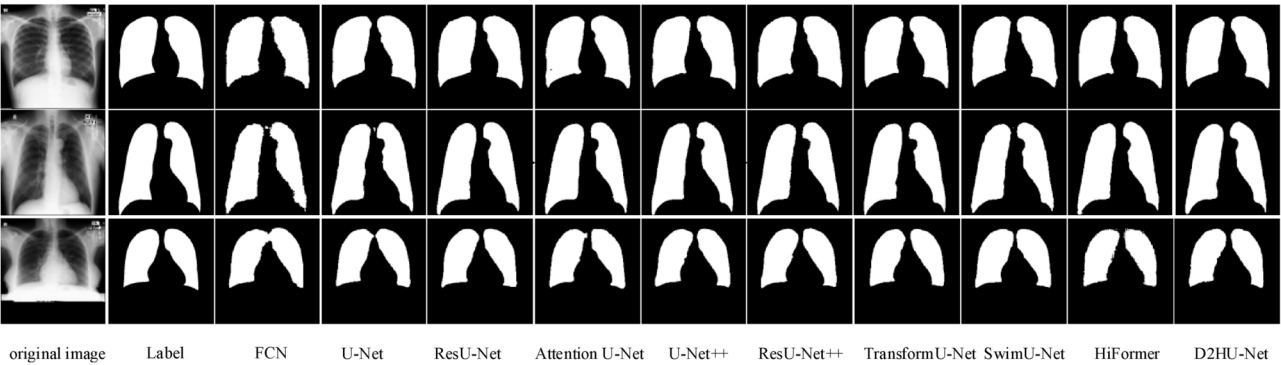


Fig. 6. Visualization of the segmentation results of D2HU-Net and 9 comparison methods on Chest X-Ray dataset.

Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
FCN	77.9 ± 0.45	67.8 ± 0.37	8.71 ± 0.42	22.57 ± 0.61	79.24 ± 3.8
U-Net	81.8 ± 0.29	73.4 ± 0.33	5.72 ± 0.71	16.41 ± 0.72	60.03 ± 1.2
ResU-Net	79.7 ± 0.15	71.1 ± 0.21	4.86 ± 0.74	18.85 ± 0.38	64.84 ± 3.5
Attention U-Net	81.4 ± 0.32	73.0 ± 0.47	8.53 ± 0.43	16.92 ± 0.45	61.28 ± 2.8
U-Net++	81.9 ± 0.44	73.6 ± 0.31	3.95 ± 0.65	15.38 ± 0.57	57.46 ± 3.9
ResU-Net++	83.3 ± 0.24	71.4 ± 0.29	-4.81 ± 0.45	23.18 ± 0.36	68.24 ± 4.1
TransformU-Net	89.2 ± 0.29	81.6 ± 0.16	3.05 ± 0.35	9.85 ± 0.28	37.36 ± 2.9
SwimU-Net	86.5 ± 0.21	78.9 ± 0.31	9.62 ± 0.29	11.2 ± 0.15	42.74 ± 3.6
HiFormer	87.2 ± 0.51	80.0 ± 0.27	4.42 ± 0.12	9.93 ± 0.11	42.44 ± 2.8
D2HU-Net	89.9 ± 0.63	81.7 ± 0.15	-11.71 ± 0.28	9.22 ± 0.09	37.34 ± 2.2

Table 4. Evaluation indicators ± standard deviation of all competing methods in the Kvasir-SEG.

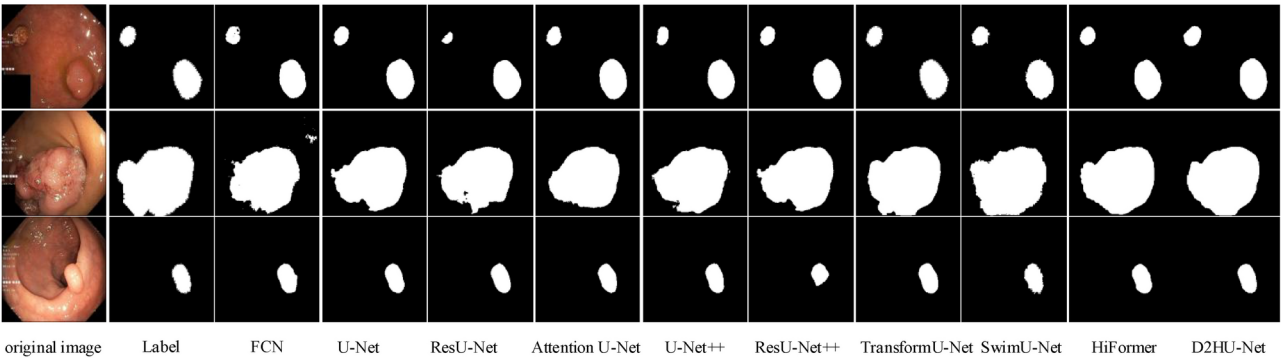


Fig. 7. Visualization of the segmentation results of D2HU-Net and 9 comparison methods on the Kvasir-SEG dataset.

Ablation studies

In the context above, we demonstrate the effectiveness of the Dynamic multi-scale layered module and multi-layer spatial attention fusion module proposed in this paper from the theoretical level. In this chapter, we will conduct a series of ablation experiments to verify the effectiveness of these modules. It includes Dynamic multi-scale layered module, multi-layer spatial attention fusion module and double branch structure. Relevant experiments were quantitatively analyzed on Synapse multi-organ segmentation dataset, Skin Lesion Segmentation dataset, Chest X-Ray Dataset and Kvasir-SEG.

1. Ablation study of dynamic multi-scale layered module: In order to study the effectiveness of dynamic multi-scale layered module, we take the traditional convolutional layer U-Net as the basis. Another contrast model is to use Dynamic multi-scale layered module instead of traditional convolutional blocks. Table 5 shows the specific index data on the four data sets, where U stands for U-Net. It can be seen that U-Net net-

Datasets	Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
Synapse multi-organ	U	75.73 ± 0.45	64.52 ± 0.52	-8.07 ± 0.39	5.83 ± 0.39	32.50 ± 3.8
	U+DML	76.31 ± 0.36	66.38 ± 0.35	4.32 ± 0.27	5.41 ± 0.42	30.47 ± 3.7
Skin Lesion	U	87.53 ± 0.47	79.83 ± 0.50	14.1 ± 0.36	13.44 ± 3.5	38.77 ± 4.2
	U+DML	88.64 ± 0.36	80.24 ± 0.45	18.34 ± 0.27	15.66 ± 4.2	51.88 ± 3.6
Chest X-Ray	U	96.4 ± 0.21	93.3 ± 0.22	1.9 ± 0.27	2.79 ± 0.44	11.08 ± 0.33
	U+DML	96.6 ± 0.15	93.8 ± 0.38	1.7 ± 0.47	2.65 ± 0.36	8.07 ± 0.45
Kvasir-SEG	U	81.8 ± 0.29	73.4 ± 0.33	5.7 ± 0.71	16.41 ± 0.72	60.03 ± 1.2
	U+DML	82.14 ± 0.48	75.64 ± 0.28	4.5 ± 0.64	14.64 ± 0.55	51.69 ± 3.5

Table 5. Ablation experimental results of dynamic multi-scale layered module.

Datasets	Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
Synapse multi-organ	D-M	79.88 ± 0.21	67.42 ± 0.22	1.94 ± 0.18	4.57 ± 0.45	25.77 ± 2.7
	D	80.39 ± 0.19	68.32 ± 0.13	1.54 ± 0.09	4.03 ± 0.36	21.49 ± 2.4
Skin Lesion	D-M	89.12 ± 0.21	81.23 ± 0.26	12.31 ± 0.07	11.24 ± 1.6	29.94 ± 1.9
	D	90.18 ± 0.17	82.01 ± 0.15	-17.9 ± 0.09	18.83 ± 1.2	50.57 ± 1.7
Chest X-Ray	D-M	97.1 ± 0.11	93.91 ± 0.19	1.1 ± 0.15	2.24 ± 0.46	6.5 ± 0.21
	D	97.2 ± 0.08	94.5 ± 0.14	3.8 ± 0.17	2.09 ± 0.2	6.08 ± 0.19
Kvasir-SEG	D-M	86.87 ± 0.66	79.75 ± 0.22	2.1 ± 0.11	12.82 ± 0.14	40.51 ± 2.4
	D	89.9 ± 0.63	81.7 ± 0.15	-11.7 ± 0.28	9.22 ± 0.09	37.34 ± 2.2

Table 6. Ablation experimental results of Multi-layer spatial attention fusion module.

Datasets	Methods	DICE (%)	IOU (%)	RAVD (%)	ASSD	MSSD
Synapse multi-organ	U	75.73 ± 0.45	64.52 ± 0.52	-8.07 ± 0.39	5.83 ± 0.39	32.50 ± 3.8
	U + DB	77.87 ± 0.28	66.82 ± 0.45	2.56 ± 0.23	4.94 ± 0.51	28.83 ± 4.1
Skin Lesion	U	87.53 ± 0.47	79.83 ± 0.50	14.1 ± 0.36	13.44 ± 3.5	38.77 ± 4.2
	U + DB	87.85 ± 0.28	79.94 ± 0.24	13.68 ± 0.14	12.55 ± 3.6	40.51 ± 4.2
Chest X-Ray	U	96.4 ± 0.21	93.3 ± 0.22	1.9 ± 0.27	2.79 ± 0.44	11.08 ± 0.33
	U + DB	96.8 ± 0.26	93.11 ± 0.27	2.4 ± 0.22	2.51 ± 0.9	7.71 ± 0.28
Kvasir-SEG	U	81.8 ± 0.29	73.4 ± 0.33	5.7 ± 0.71	16.41 ± 0.72	60.03 ± 1.2
	U + DB	84.31 ± 0.97	76.66 ± 0.48	3.6 ± 0.45	15.76 ± 0.34	49.68 ± 2.7

Table 7. Results of ablation experiments for dual decoding paths.

work is superior to U-Net in all indicators after adding DML. IOU improved by 1.82% on the Synapse multi-organ segmentation dataset. On Skin Lesion Segmentation, DICE improved by 1.11%. In the Chest X-Ray dataset, due to the simplicity of the dataset, the values of each indicator were relatively high, but there were small improvements in the indicator data. On Kvasir-SEG, JC is up 2.2 percent. It can be seen from various data that Dynamic multi-scale layered module is helpful to improve the accuracy of segmentation.

2. Ablation Study of multi-layer spatial attention fusion module: We first tested the D2HU-Net proposed in this paper on four data sets. Second. We removed MSAF from D2HU-Net, and then conducted training on four data sets before testing. Table 6 shows the specific indicators of D2HU-Net on four data sets, Where D stands for D2HU-Net and M stands for MSAF. In the Synapse multi-organ segmentation dataset, compared with D2HU-Net without MSAF, DICE of D2HU-Net increased by 0.51% and IOU increased by 1.86%. RAVD decreased by 3.75, ASSD and MSSD decreased by 0.42 and 2.03, respectively. In terms of Skin Lesion Segmentation, compared with D2HU-Net without MSAF, DICE of D2HU-Net increased by 1.06% and IOU increased by 0.78%. On the Chest X-Ray dataset, DICE was increased by 0.1%. IOU increased by 0.59%. On Kvasir-SEG, IOU of D2HU-Net increased by 3.07% and IOU increased by 1.95% compared with D2HU-Net without MSAF. It can be seen from the evaluation indexes of the four data sets that the MASF module proposed in this paper is reasonable and helpful to improve the segmentation accuracy of the model.
3. Ablation Study of Double branch: In order to verify the rationality of the double branch structure proposed in this paper, we chose U-Net as the base model, and then added double branch structure to the U-Net network model. The model was evaluated on four data sets. Table 7 shows the evaluation index data on the four datasets, where U stands for U-Net and DB stands for double branch. It can be seen from the data in the table that, combined with the two-branch U-Net network, all evaluation indexes are better than the benchmark

Methods	GFLOPs	Inference speed (ms)	Paras (M)
FCN	19.52	22.84	18.6
U-Net	40.23	17.78	17.27
ResU-Net	68.08	32.34	13.04
Attention U-Net	30.91	29.77	34.88
U-Net++	34.63	24.49	9.16
ResU-Net++	70.84	36.08	14.48
TransformU-Net	78.63	41.21	105.28
SwimU-Net	40.52	22.34	27.17
HiFormer	45.72	28.31	25.51
D2HU-Net	32.94	21.13	41.58

Table 8. Comparison of computational complexity.

model U-Net, which can also be seen that the first decoding branch has a guiding effect on the second decoding branch, which improves the accuracy of medical image segmentation.

Complexity calculation

In addition to segmentation performance, factors such as inference speed and model size play a pivotal role in determining the practical applicability of clinical diagnostics. The computational complexity of a model can be quantified using the well-known gigafloating-point operations (GFLOPs), inference speed, and the number of parameters. As illustrated in Table 8, D2HU-Net demonstrates notable strengths, ranking third in terms of GFLOPs and second in terms of inference speed. The inference speed is a crucial metric in clinical environments, especially when dealing with a large number of medical images, and D2HU-Net demonstrates superior performance in this regard. However, it should be noted that D2HU-Net requires more parameters due to the presence of two decoding paths, i.e., the dynamic multiscale layering module and the multilayer spatial attention fusion module. Despite its augmented complexity, D2HU-Net exhibits superiority in medical image segmentation accuracy, a trade-off that involves sacrificing spatial complexity for enhanced segmentation precision.

Discussion

Our comprehensive experiments on four different modal medical image datasets show that D2HU-Net has certain advantages over other SOTA. In quantitative analysis, all the data show that D2HU-Net can perform multi-modal medical image segmentation very well. From the qualitative analysis, D2HU-Net can also be accurately segmented into target regions, which is also the same as the result of quantitative analysis. In contrast, our D2HU-Net still has some challenges in processing low-contrast, borderless medical images, resulting in inaccurate segmentation. In future work, we will investigate more robust segmentation methods.

Conclusion

In this paper, we propose a medical image segmentation method based on D2HU-Net network, which has three main innovative contributions. Firstly, the proposed D2HU-Net network adopts a double-branch decoding path, the shallow decoding path provides guidance for the deep decoding path, and improves the generalization ability of the network. Secondly, a dynamic multi-scale hierarchical structure is proposed, which can selectively utilize semantic information at different scales, integrate multi-scale features, and refine the features at each scale. The multi-scale expression of network features is enhanced. Thirdly, a multi-layer spatial attention fusion module is proposed. It not only focuses on the distinguishing features of channel and spatial dimension, but also establishes the multi-dimensional interactive relationship between channel and spatial dimension. In addition, due to the different importance of each branch feature map, we propose an adaptive weighting coefficient for each branch to more efficiently fuse feature maps from all four branches. Finally, selective nuclear feature aggregation can dynamically aggregate advanced features of different branches/types, effectively improving the accuracy of medical image segmentation.

Experimental results on four different modal medical image datasets demonstrate the effectiveness and robustness of the proposed D2HU-Net network. At the same time, a series of ablation experiments verify the effectiveness of each module in the proposed D2HU-Net network. Although the proposed D2HU-Net has achieved good classification performance, there is still room for improvement. For example, the proposed dynamic multi-scale hierarchical structure can extract multi-scale features, increase the number of parameters in the network, and improve the complexity of the model. While the experiments confirmed the robustness of D2HU-Net on various datasets, there is still potential to improve the model's resilience to extremely noisy data or images under extreme conditions. In further work, we will explore how to make the network more lightweight and highly robust to adapt to more segmentation scenarios while improving the accuracy of the network.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 25 September 2024; Accepted: 3 March 2025

Published online: 10 March 2025

References

- Radak, M., Lafta, H. Y. & Fallahi, H. Machine learning and deep learning techniques for breast cancer diagnosis and classification: a comprehensive review of medical imaging studies. *J. Cancer Res. Clin. Oncol.* **149**, 10473–10491 (2023).
- Qureshi, I. et al. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **90**, 316–352 (2023).
- 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).
- Zhang, Y. et al. Interactive medical image annotation using improved attention u-net with compound geodesic distance. *Expert Syst. Appl.* **237**, 121282 (2024).
- Siyi, X. et al. Arga-unet: Advanced u-net segmentation model using residual grouped convolution and attention mechanism for brain tumor mri image segmentation. *Virtual Real. Intell. Hardw.* **6**, 203–216 (2024).
- Wang, H., Cao, P., Wang, J. & Zaiane, O. R. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. *Proc. AAAI Conf. Artif. Intell.* **36**, 2441–2449 (2022).
- Zunair, H. & Hamza, A. B. Sharp u-net: Depthwise convolutional network for biomedical image segmentation. *Comput. Biol. Med.* **136**, 104699 (2021).
- Li, Z., Liu, F., Yang, W., Peng, S. & Zhou, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **33**, 6999–7019 (2021).
- Agarwal, R. et al. Deep quasi-recurrent self-attention with dual encoder-decoder in biomedical ct image segmentation. *IEEE J. Biomed. Health Inf.* (2024).
- Phan, T.-D.-T., Kim, S.-H., Yang, H.-J. & Lee, G.-S. Skin lesion segmentation by u-net with adaptive skip connection and structural awareness. *Appl. Sci.* **11**, 4528 (2021).
- Chen, B., Liu, Y., Zhang, Z., Lu, G. & Kong, A. W. K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Trans. Emerg. Top. Comput. Intell.* (2023).
- Huang, Z., Wang, Z., Yang, Z. & Gu, L. Adwu-net: adaptive depth and width u-net for medical image segmentation by differentiable neural architecture search. In *International Conference on Medical Imaging with Deep Learning*, 576–589 (PMLR, 2022).
- Jabeen, K. et al. A novel fusion framework of deep bottleneck residual convolutional neural network for breast cancer classification from mammogram images. *Front. Oncol.* **14**, 1347856 (2024).
- Senapati, P., Basu, A., Deb, M. & Dhal, K. G. Sharp dense u-net: an enhanced dense u-net architecture for nucleus segmentation. *Int. J. Mach. Learn. Cybern.* **15**, 2079–2094 (2024).
- Hu, Y., Deng, L., Wu, Y., Yao, M. & Li, G. Advancing spiking neural networks toward deep residual learning. *IEEE Trans. Neural Netw. Learn. Syst.* (2024).
- Zuo, Q., Chen, S. & Wang, Z. R2au-net: attention recurrent residual convolutional neural network for multimodal medical image segmentation. *Secur. Commun. Netw.* **2021**, 6625688 (2021).
- Fu, J. et al. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154 (2019).
- Sun, G. et al. Da-transunet: integrating spatial and channel dual attention with transformer u-net for medical image segmentation. *Front. Bioeng. Biotechnol.* **12**, 1398237 (2024).
- Han, K. et al. Transformer in transformer. *Adv. Neural. Inf. Process. Syst.* **34**, 15908–15919 (2021).
- Agarwal, R., Ghosal, P., Sadhu, A. K., Murmu, N. & Nandi, D. Multi-scale dual-channel feature embedding decoder for biomedical image segmentation. *Comput. Methods Programs Biomed.* **257**, 108464 (2024).
- Kussul, N., Lavreniuk, M., Skakun, S. & Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **14**, 778–782 (2017).
- Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440 (2015).
- Xiao, X., Lian, S., Luo, Z. & Li, S. Weighted res-unet for high-quality retina vessel segmentation. In *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, 327–331 (IEEE, 2018).
- Radiuk, P. Applying 3d u-net architecture to the task of multi-organ segmentation in computed tomography. *Appl. Comput. Syst.* **25**, 43–50 (2020).
- Zhao, H., Qi, X., Shen, X., Shi, J. & Jia, J. Icnnet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 405–420 (2018).
- Chen, W., Jiang, Z., Wang, Z., Cui, K. & Qian, X. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8924–8933 (2019).
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 834–848 (2017).
- Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803 (2018).
- Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19 (2018).
- Fu, J. et al. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3146–3154 (2019).
- Wang, G. et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE Trans. Med. Imaging* **37**, 1562–1573 (2018).
- Hu, J., Shen, L. & Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7132–7141 (2018).
- Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3–19 (2018).
- Chen, B., Liu, Y., Zhang, Z., Lu, G. & Kong, A. W. K. Transattunet: Multi-level attention-guided u-net with transformer for medical image segmentation. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2023).
- Huang, Z. et al. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 603–612 (2019).
- Misra, D., Nalamada, T., Arasanipalai, A. U. & Hou, Q. Rotate to attend: Convolutional triplet attention module. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3139–3148 (2021).
- Jaeger, S. et al. Automatic tuberculosis screening using chest radiographs. *IEEE Trans. Med. Imaging* **33**, 233–245 (2013).
- Mou, L., Hua, Y. & Zhu, X. X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12416–12425 (2019).

39. Diakogiannis, F. I., Waldner, F., Caccetta, P. & Wu, C. Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS J. Photogramm. Remote. Sens.* **162**, 94–114 (2020).
40. Lian, S. et al. Attention guided u-net for accurate iris segmentation. *J. Vis. Commun. Image Represent.* **56**, 296–304 (2018).
41. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, 3–11 (Springer, 2018).
42. Jha, D. et al. Resunet++: An advanced architecture for medical image segmentation. In *2019 IEEE International Symposium on Multimedia (ISM)*, 225–2255 (IEEE, 2019).
43. Cao, H. et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*, 205–218 (Springer, 2022).
44. Chen, J. et al. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306* (2021).
45. Heidari, M. et al. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6202–6212 (2023).

Acknowledgements

This work was supported by the Scientific Research Fund Project of Qiqihar Academy of Medical Sciences (QM-SI2022M-05).

Author contributions

Z.W. and S.F. made significant contributions to the conceptualization and design of this study, as well as to the analysis of the data. H.Z., C.W., C.X., P.H., C.S., and G.S. made significant contributions to the acquisition of the data. All authors have approved the submitted version and any substantially modified version involving their contributions to the study. All authors have consented to assume responsibility for their contributions and to ensure that issues related to the accuracy or completeness of any part of the work, even if not directly related to the authors themselves, are properly investigated, resolved, and documented in the literature.

Declarations

Competing interests

The authors declare no competing interests

Additional information

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

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025