Original Article

Automatic computed tomography image segmentation method for liver tumor based on a modified tokenized multilayer perceptron and attention mechanism

Bo Yang[1], **Jie Zhang**[2], **Youlong Lyu**[2], **Jun Zhang**[3]

[1]College of Mechanical Engineering, Donghua University, Shanghai, China; [2]Institute of Artificial Intelligence, Donghua University, Shanghai, China; [3]College of Information Science and Technology, Donghua University, Shanghai, China

Contributions: (I) Conception and design: B Yang; (II) Administrative support: B Yang, Jie Zhang; (III) Provision of study materials or patients: B Yang; (IV) Collection and assembly of data: B Yang, Jun Zhang; (V) Data analysis and interpretation: B Yang, Y Lyu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Youlong Lyu, PhD. Institute of Artificial Intelligence, Donghua University, 2999 Renmin North Road, Songjiang District, Shanghai 201620, China. Email: lvyoulong@dhu.edu.cn.

Background: The automatic medical image segmentation of liver and tumor plays a pivotal role in the clinical diagnosis of liver diseases. A number of effective methods based on deep neural networks, including convolutional neural networks (CNNs) and vision transformer (ViT) have been developed. However, these networks primarily focus on enhancing segmentation accuracy while often overlooking the segmentation speed, which is vital for rapid diagnosis in clinical settings. Therefore, we aimed to develop an automatic computed tomography (CT) image segmentation method for liver tumors that reduces inference time while maintaining accuracy, as rigorously validated through experimental studies.

Methods: We developed a U-shaped network enhanced by a multiscale attention module and attention gates, aimed at efficient CT image segmentation of liver tumors. In this network, a modified tokenized multilayer perceptron (MLP) block is first leveraged to reduce the feature dimensions and facilitate information interaction between adjacent patches so that the network can learn the key features of tumors with less computational complexity. Second, attention gates are added into the skip connections between the encoder and decoder, emphasizing feature expression in relevant regions and enabling the network to focus more on liver tumor features. Finally, a multiscale attention mechanism autonomously adjusts weights for each scale, allowing the network to adapt effectively to varying sizes of liver tumors. Our methodology was validated via the Liver Tumor Segmentation 2017 (LiTS17) public dataset. The data from this database are from seven global clinical sites. All data are anonymized, and the images have been prescreened to ensure the absence of personal identifiers. Standard metrics were used to evaluate the performance of the model.

Results: The 21 cases were included for testing. The proposed network attained a Dice score of 0.713 [95% confidence interval (CI): 0.592–0.834], a volumetric overlap error of 0.39 (95% CI: 0.17–0.61), a relative volume difference score of 0.19 (95% CI: −0.37 to 0.31), an average symmetric surface distance of 2.04 mm (95% CI: 0.89–4.19), a maximum surface distance of 9.42 mm (95% CI: 6.97–19.87), and an inference time of 26 ms on average for liver tumor segmentation.

Conclusions: The proposed network demonstrated efficient liver tumor segmentation performance with less inference time. Our findings contribute to the application of neural networks in rapid clinical diagnosis and treatment.

Keywords: Liver tumor segmentation; multilayer perceptron-based network (MLP-based network); tokenized MLP; multiscale attention structure; attention gate

Submitted Oct 03, 2024. Accepted for publication Jan 23, 2025. Published online Feb 26, 2025.
doi: 10.21037/qims-24-2132
View this article at: https://dx.doi.org/10.21037/qims-24-2132

Introduction

The liver, a critical organ in human metabolic function and the vascular system, features a complex anatomical structure (1). Malignant tumor of the liver is a major threat to human health, and its incidence rate is increasing each year (2). Clinicians typically arrive at a preliminary diagnosis of liver tumors through segmenting diseased tissues via computed tomography (CT) images to evaluate the size, shape, and location of tumors (3). In clinical practice, tumor tissue segmentation is typically sketched manually by expert radiologists, which is tedious and time-consuming. Moreover, the accuracy of the segmentation results depends highly on the experience and skill of the operators (4). With the growing preference for real-time intraoperative image processing and the development of mobile medical diagnostic equipment, there is a substantial demand for rapid medical image processing (5-7). Therefore, developing an accurate and efficient automatic liver tumor segmentation method may provide considerable value for the clinical diagnosis and treatment of liver disease.

However, automated liver tumor segmentation from CT scans involves several challenges. First, the boundary between liver tumor and normal liver tissue may be unclear (8). Second, the shape, location, and volume of liver tumors vary from patient to patient, complicating accurate segmentation (9). Third, medical image data impose a considerable memory burden, and many of the current image segmentation algorithms involve high computational complexity, thus consuming a substantial amount of computing resources (10).

Traditional CT image segmentation techniques, such as thresholding (11), region growing (12), and clustering-based methods (13), struggle with the indistinct boundaries between liver tumors and normal tissues and fail to effectively segment multiscale liver tumors. With the proliferation of deep learning technologies (14), a number of deep neural networks with automatic feature extraction and strong learning ability have been developed to segment liver tumors. Convolutional neural network (CNN)-based methods (15-19) and vision transformer (ViT)-based methods (20-23), due to their encoder-decoder structures with skip connections, are the foundation of medical image segmentation and form the backbone of almost all leading neural networks in medical image segmentation. In recent years, various CNN- and ViT-based networks have been proposed for the accurate segmentation of liver tumors (24,25). Liu *et al.* (26) introduced a spatial feature fusion convolutional network designed for segmenting liver tissue and liver tumor from CT images, extracting side-outputs at each convolutional block to effectively harness multiscale features. Luan *et al.* (27) designed a CNN-based network that adhered to the classical encoder-decoder configuration, incorporating long-hop connections across layers in both the contraction and the expansion path to merge the semantic information. Meng *et al.* (28) proposed a biphasic CNN algorithm for liver tumor segmentation. In this algorithm, the first stage involves localizing the liver region by using the spatial information of shallow features, and the second stage segments tumors from the segmentation region of the first stage. Cheng *et al.* (29) devised an edge-guide segmentation network based on transformer that captures richer contextual information and integrates multi-scale features, boosting liver tumor segmentation performance. Li *et al.* (30) designed a dynamic hierarchical transformer architecture that augments the semantic representation of tumor features through use of hierarchical operations across varying receptive fields. Di *et al.* (31) created an innovative hybrid end-to-end network that amalgamates transformers and directional data to extract multilevel features. However, the computational complexity of CNN-based networks increases with the number of network layers, whereas that of ViT-based networks increases due to self-attention mechanisms. These networks are characterized by a plenitude of parameters, which reduce the inference speed and limit their broader application in rapid clinical diagnosis and real-time intraoperative image processing.

To enhance inference speed, the multilayer perceptron (MLP)-mixer (MLP-mixer) architecture (32) has been proposed, which employs the linear complexity of MLP layers to reduce parameters while effectively extracting and integrating the diverse features of the target. However, the MLP-mixer struggles with the variability in input image sizes, posing challenges in the segmentation of liver tumors of different sizes. Subsequently, a tokenized MLP block (5), based on the MLP-mixer architecture, was designed to enhance the segmentation performance through implementation of a window-based attention mechanism. Although these MLP-based networks, employing advanced structures of CNNs and ViTs, have achieved good performance with minimal complexity, they face challenges in learning sufficient information and features, leading to limitations in the segmentation performance of liver tumors.

Attention mechanisms are being increasingly embedded within networks to boost segmentation performance. These include spatial attention mechanisms (33,34), channel attention mechanisms (35,36), and mixed attention mechanisms (37,38), among others. The spatial attention mechanism enhances the network's ability to capture local feature information by allocating different attention weights across the feature space. Meanwhile, the channel attention mechanism aims to automatically identify and prioritize the essential information of each channel, thus allocating greater attention to specific channels. The mixed attention mechanism captures more contextual information by integrating multiple attention mechanisms. Numerous researchers have proposed segmentation networks for medical images based on these attention mechanisms. Guo *et al.* (34) designed a spatial attention mechanism for retinal vessel segmentation, which processes feature maps across spatial dimensions and adjusts the input feature map for refined feature representation. Chen *et al.* (39) introduced a multilevel attention mechanism that enhances consistency in feature representation and semantic embeddings by a acquiring greater amount of contextual semantic information and data on global spatial relationships. Zhou *et al.* (40) described a method that employs both channel and spatial attention mechanisms to adjust feature representation spatially and in channel-wise fashion for automatic coronavirus disease 2019 (COVID-19) CT segmentation. Overall, these attention mechanisms allow networks to assimilate a greater degree of detail from the information of targets by discerning key aspects of both spatial and channel dimensions. However, they have higher computational complexity and exhibit limited adaptability to variability in the size and characteristics of liver tumors.

In this paper, we propose a two-stage network that combines the CNN stage with the modified tokenized MLP stage, following the U-shaped structure with attention gates and a multiscale attention module, aimed at achieving high segmentation performance of liver tumors with fewer parameters. Specifically, we follow the network design of CNNs in the shallow layers and MLPs in the deep layers. Additionally, the spatial-shift operation enhances the information transfer between adjacent patches with minimal computational overhead. Subsequently, attention gates, integrated within the skip connections, enhance the other key features of liver tumors. A multiscale attention mechanism then combines the features from different layers, which assists the network in the processing of liver tumors of different sizes. It is worth mentioning that attention gates and the multiscale attention mechanism minimally increase the number of network parameters. To verify the liver tumor segmentation performance of the proposed model, comparative and ablation studies were conducted. The main contributions of this paper are as follows:

(I) A modified tokenized MLP module is proposed, which used spatial-shift operations to shift feature maps across different patches, facilitating information interaction between adjacent patches with less computational complexity.

(II) We describe the integration of attention gate mechanisms within the skip connections of the U-shaped network, which suppresses the learning of irrelevant region features in the image to address segmentation challenges arising from blurred boundaries between liver tumors and normal tissues.

(III) A multiscale attention mechanism is introduced, which assigns weights into different feature maps at multiple scales to learn the correlation between information of different layers and the segmentation object, achieving precise segmentation of liver tumors of various sizes.

We present this article in accordance with the TRIPOD + AI reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-24-2132/rc).

Methods

The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

This Methods includes four subsections. First, the patient data are described. Second, the proposed methodologies are elaborated upon. Third, the implementation details are outlined. Finally, the evaluation metrics are identified.

Patient data

The Liver Tumor Segmentation 2017 (LiTS17) dataset (4) is a CT dataset specifically designed for the segmentation of the liver and liver tumors. It includes both primary and secondary tumors, presenting a diverse range of sizes and appearances, as well as varying lesion-to-background densities (hyper- and hypodense). The CT images of 131 patients in the LiTS2017 dataset are randomly split into three sets: the training set, the validation set, and test set. The training and the validation sets contain 90 and 21 images, respectively, while the remaining images are used as the test set. All the medical images have a resolution of 512×512 in resolution. The number of slices ranges from 42 to 1,026, and the spacing between slices ranges from 0.45 to 6.0 mm. For this study, the images were preprocessed by the methods described in the preprocessing section.

The data were collected from seven clinical sites all over the world, including (I) Rechts der Isar Hospital, the Technical University of Munich (Germany), (II) Radboud University Medical Center (the Netherlands), (III) Polytechnique Montréal and CHUM Research Center (Canada); (IV) Sheba Medical Center (Israel); (V) the Hebrew University of Jerusalem (Israel); (VI) Hadassah University Medical Center (Israel); and (VII) the Institute for Research into Cancer of the Digestive System (IRCAD; France). All data were anonymized, and images were rigorously inspected to preclude the presence of personal identifiers.

The dataset has been manually labeled for liver tissue and liver tumor regions by four radiologists from different institutions. In our study, the labeled data in the test set served as the ground truth and were used to assess the segmentation performance of the model.

Proposed methods

In this section, the medical image preprocessing step, before data are sent into the network, is introduced, and then the proposed network and designs are described in detail.

Preprocessing

Most semantic segmentation networks are two-dimensional (2D). However, many medical images, such as those of magnetic resonance imaging and CT, exist in three-dimensional (3D) forms. The 3D medical images contain a series of continuous 2D slices, requiring a considerable amount of training time and a complex hardware configuration (41). The models trained by 3D medical images may be parameter-heavy, resulting in a large amount of inference time. To reduce the training time and hardware requirements, some researchers have transformed 3D datasets into 2D datasets, which causes a loss of intraslice information and reduces segmentation accuracy due to their being a limited receptive filed (42). In this study, 2.5-dimentional (43,44) datasets were incorporated into our network. As can be seen in *Figure 1*, 3D medical images are cut into several stacks with three 2D slices, which reduces the dataset size and retains a sufficient amount of intraslice information.

Proposed network

The proposed network (see *Figure 2*) consists of four main components: a convolutional stage, an MLP stage, attention gates, and a multiscale attention mechanism. The whole network architecture follows a U-shape with a skip connection, comprising two main sections: the encoder and the decoder. The input image is passed through the encoder where the first three blocks are convolutional blocks and the next two are tokenized MLP blocks. The decoder contains two modified tokenized MLP blocks and three convolutional blocks. Every decoder block has a skip connection from a parallel encoder block, and attention gates are added into every skip connection. At the end of the output layer, the multiscale attention structure recovers the input image. The channels' number at each stage is as follows: C1=32, C2=64, C3=128, C4=160, and C5=256.

(I) Convolutional stage: purely MLP-based network architectures, such as MLP-mixer (32), have limitations in capturing local spatial details. To extract more features of liver tumors with fewer parameters, a single-layer CNN was employed. In this approach, every convolutional block contains a convolution layer, a batch normalization layer, and a rectified linear unit (ReLU) activation layer. The kernel size of each convolution layer is 3×3, and the parameters of stride and padding are 1 respectively. A max pooling layer with a pool window 2×2 is used between every stage in the encoder section, while a bilinear interpolation layer is applied to upsample the feature maps in the decoder section.

(II) MLP stage: the tokenized MLP block (5) has demonstrated desirable performance in medical image segmentation. However, this approach focuses solely on feature extraction within tokens while overlooking the information exchange between tokens. To facilitate the network's interaction of intertoken information, spatial shift

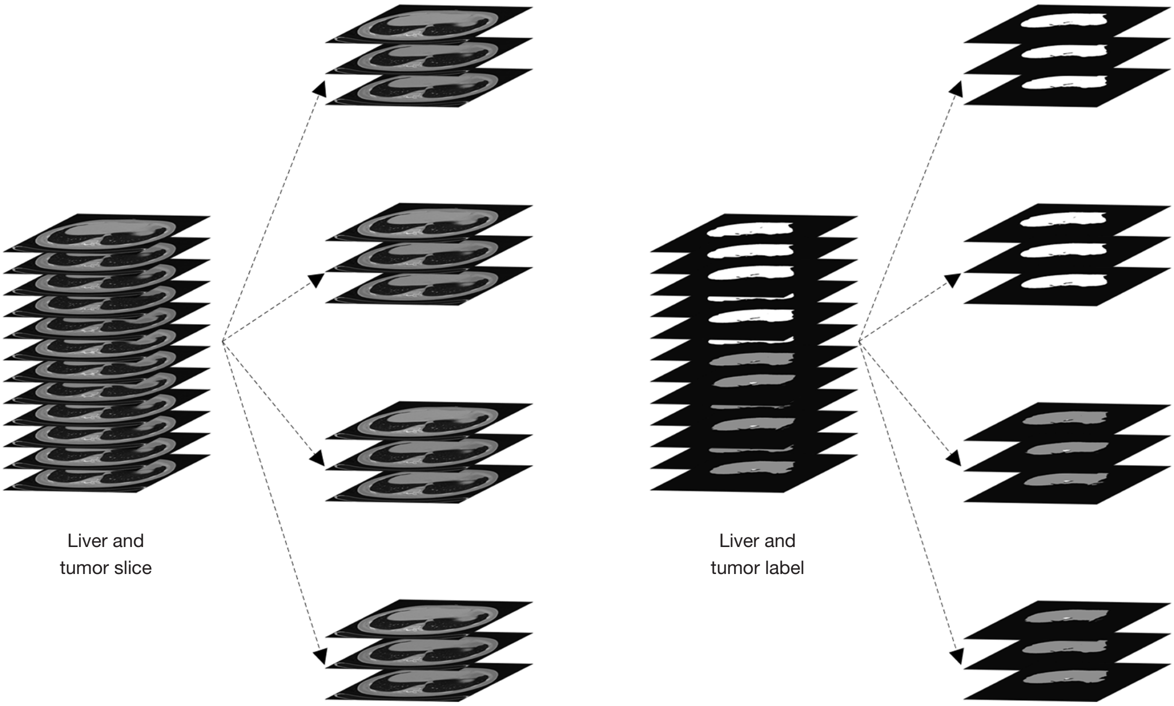

Figure 1 Preprocessing steps for the training dataset.

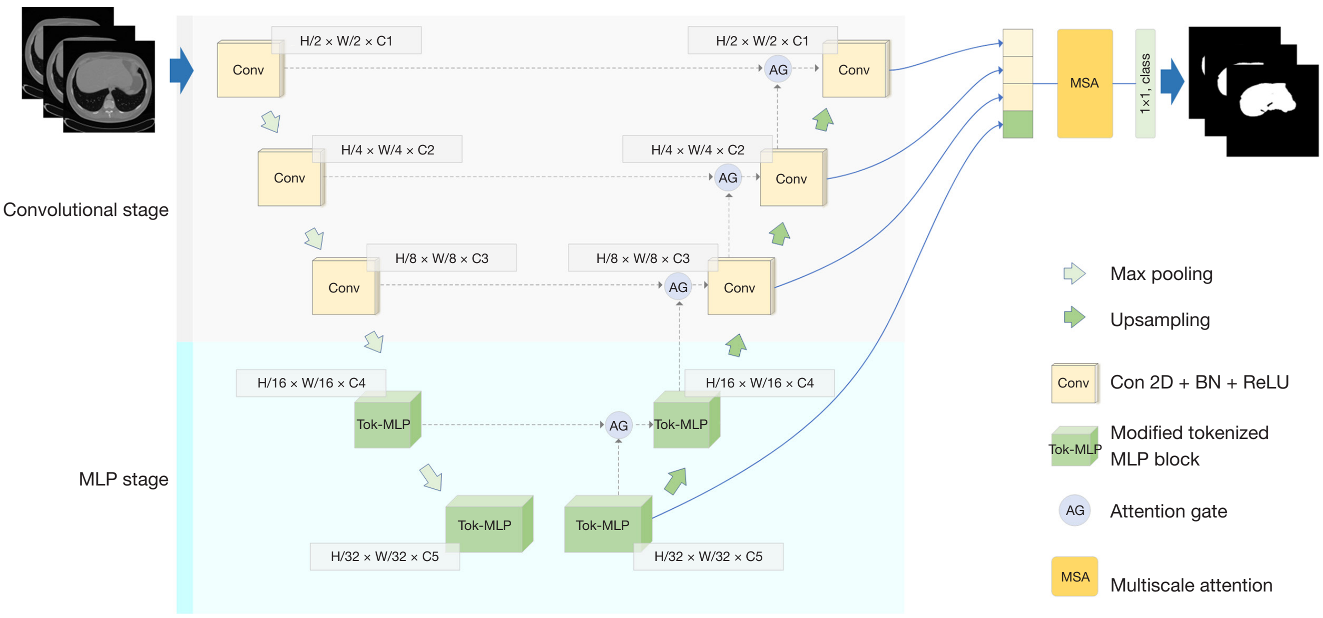

Figure 2 The proposed network architecture. BN, batch normalization; ReLU, rectified linear unit; MLP, multilayer perceptron; Conv, convolution; MSA, multiscale attention.

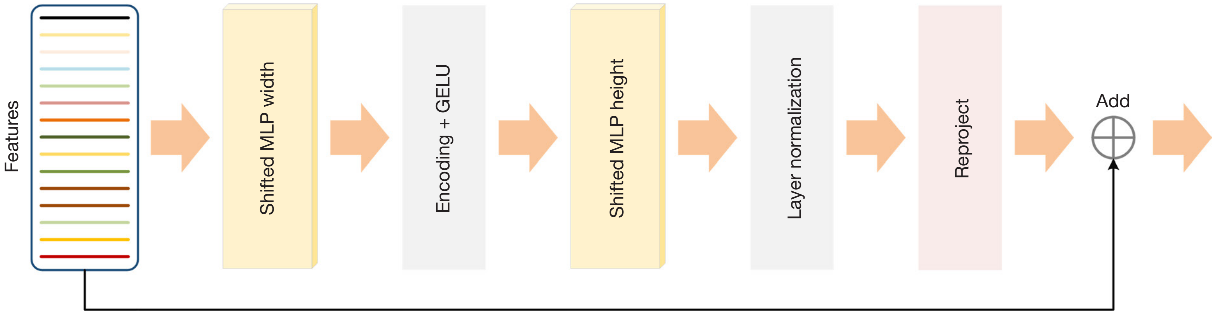

Figure 3 The modified tokenized MLP block. GELU, Gaussian error linear unit; MLP, multilayer perceptron.

operations (45) were introduced into the tokenized MLP, offering improved segmentation performance at a lower computational cost.

(i) Modified tokenized MLP block. As shown in *Figure 3*, each modified tokenized MLP block is equipped with a shifted MLP width layer, Gaussian error linear unit (GELU) activation, an encoding layer, a shifted MLP height layer, a normalization layer, and a reproject layer. First, the features are sent into a modified shifted MLP (across width), where the features are shifted and projected into tokens. The tokens are then sent into the MLP, after which the features are passed through a depth-wise convolutional layer that encodes positional information of the MLP features, thereby reducing the number of parameters. A GELU function is used to be as an activation layer instead of a ReLU for better performance. Next, the features are passed through the another modified shifted MLP (across height), which is followed by layer normalization. In the end, all the features are reprojected to the initial state, and a residual connection here is used to add the original tokens. The computational processes in the modified tokenized MLP block can be summarized as follows:

$$X_{shift} = Shift_W(X), T_W = Tokenize(X_{shift}) \quad [1]$$

$$Y = f\{DWConv[MLP(T_W)]\} \quad [2]$$

$$Y_{shift} = Shift_H(Y), T_H = Tokenize(Y_{shift}) \quad [3]$$

$$Y = f(LN(T + MLP(GELU(T_H)))) \quad [4]$$

where W denotes width, H denotes height, T denotes the tokens, $DWConv$ denotes depth-wise convolution, and LN denotes layer normalization.

(ii) Shifted MLP. As depicted in *Figure 4*, the features are first split into several partitions and then shifted sequentially across height and width before tokenization operation, which help the network learn local and global features. Finally, the tokenized features are sent into the MLP.

(iii) Spatial shift. As can be seen in *Figure 5*, the input of the spatial shift block is denoted as $X \in \mathbb{R}^{h \times w \times c}$, where h is the height, w is the width, and c is the number of channels. The spatial shift operation is executed in two steps. First, X is split to k thinner tensors $\{X_\tau\}_{\tau=1}^{k}$, where $X_\tau \in \mathbb{R}^{h \times w \times c/k}$, and k is dependent on the design of the shifting directions in the second step. Shifting along four directions was found to be sufficient to enable the necessary information interaction between different tokens, and thus we set k to 4. Second, each group is moved in different directions. The first group X_1 is shifted along the width dimension by +1. In parallel, the second group X_2 is shifted along the wide dimension by −1. Similarly, the third group X_3 is shifted upward along the height dimension by +1, and the final group X_4 is shifted along the height dimension by −1. After spatial shift operation, each token absorbs the feature information from its adjoining patches. Thus, the information interaction between different tokens is enabled.

(III) Attention gate: the attention gate mechanism,

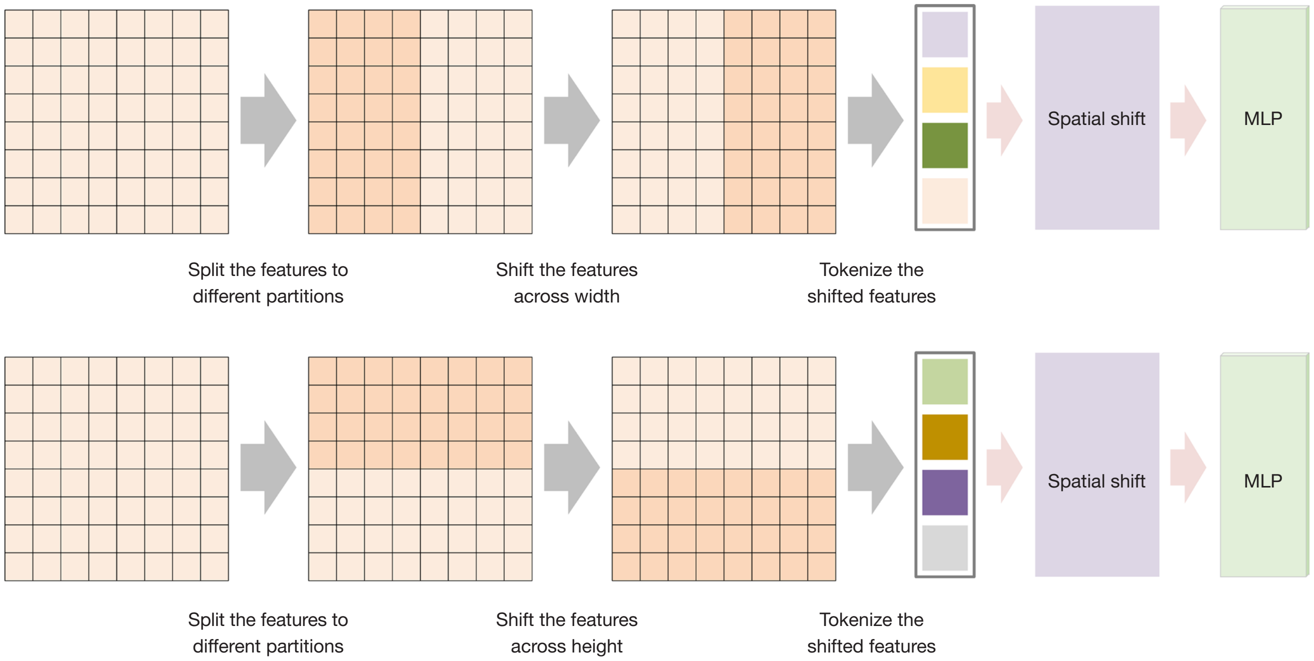
Spatial shift
MLP
Split the features to different partitions
Shift the features across width
Tokenize the shifted features
Spatial shift
MLP
Split the features to different partitions
Shift the features across height
Tokenize the shifted features

Figure 4 The shifted MLP block schematic. MLP, multilayer perceptron.

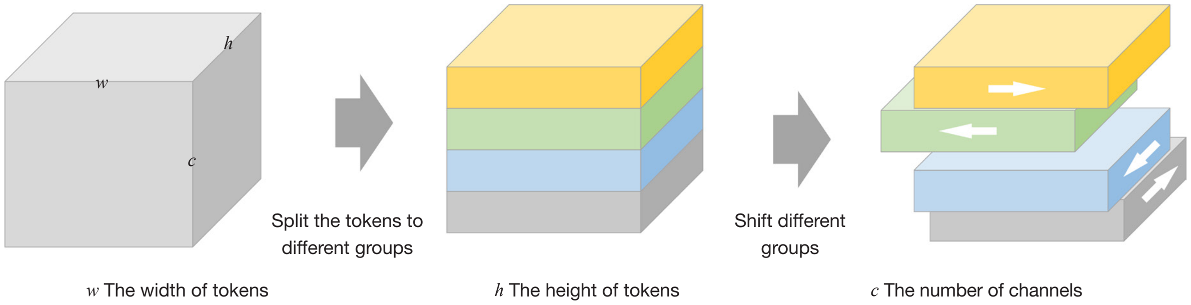

Figure 5 Schematic diagram of the spatial shift block.

originating from Attention-UNet (46), implicitly learns to suppress irrelevant areas in an input image while accentuating significant features useful for specific tasks, with minimal computational overhead. To have the proposed network place greater focus on the features of liver tumors, the attention gate mechanism is embedded into our segmentation network. The structure of the attention gate is shown in *Figure 6*, where g is the feature from the decoder section, and x is the feature from the encoder section. Parallel operations are performed on g and x, followed by a ReLU activation function and a 1×1 convolution. Subsequently, the sigmoid function is used to activate the feature maps, and they are resampled to obtain the attention weight coefficient α. Finally, $\hat{x}$ is obtained by $\alpha \times x$.

(IV) Multiscale attention structure: liver tumors have different shapes and sizes, complicating the segmentation process for general medical segmentation networks. For obtaining accurate tumor segmentation results, a rational approach is to combine different layers' features for the final prediction. However, the layers have different scales

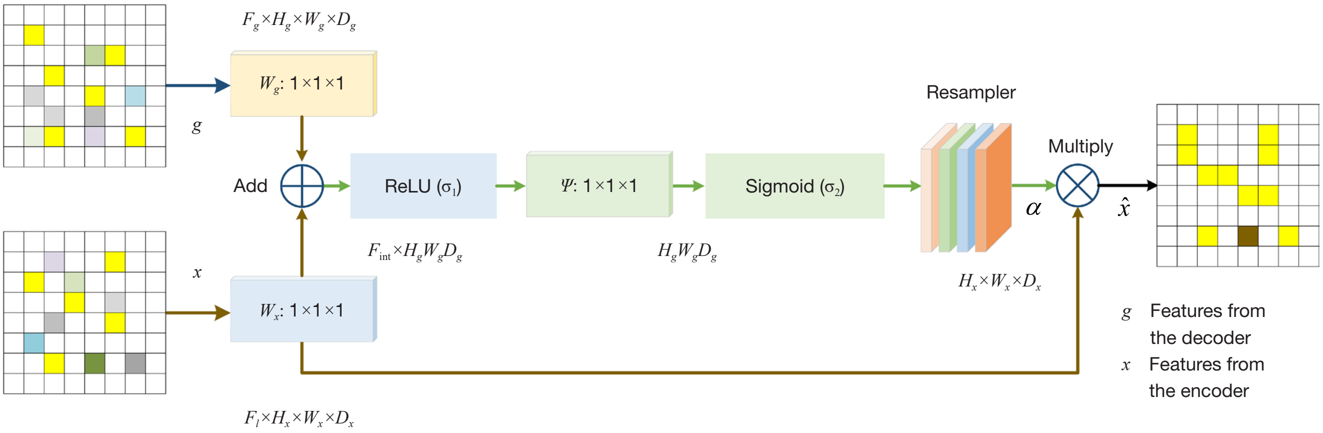
$F_g \times H_g \times W_g \times D_g$
W_g: 1×1×1
g
Add
ReLU (σ_1)
Ψ: 1×1×1
Sigmoid (σ_2)
Resampler
Multiply
α
$\hat{x}$
x
W_x: 1×1×1
$F_{int} \times H_g W_g D_g$
$H_g W_g D_g$
$H_x \times W_x \times D_x$
$F_l \times H_x \times W_x \times D_x$
g Features from the decoder
x Features from the encoder

Figure 6 The attention gate structure. ReLU, rectified linear unit.

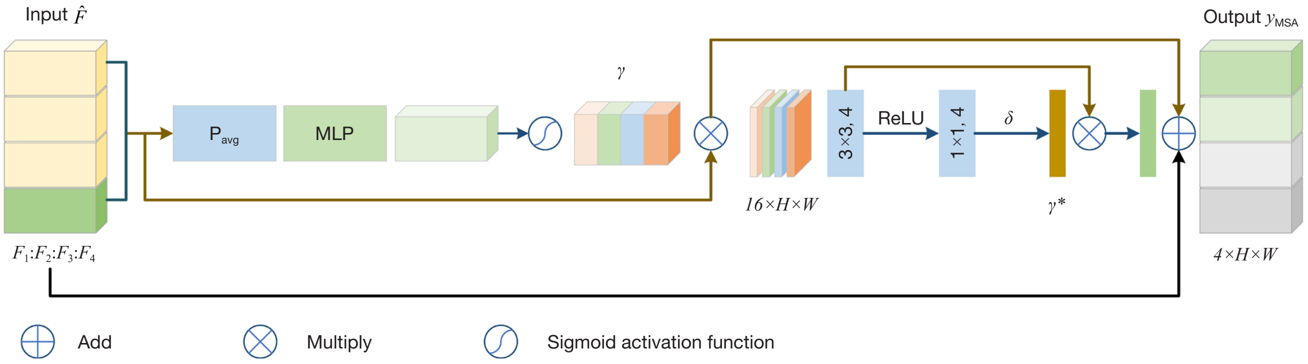

Figure 7 The multiscale attention structure. P_{avg}, average pooling; MLP, multilayer perceptron; ReLU, rectified linear unit; MSA, multiscale attention.

feature maps, whose relevance to the input varies. To allow the network to learn the corresponding scales of the input, a multiscale attention structure was developed.

The proposed multiscale attention structure is illustrated in *Figure 7*. In this structure, the bilinear interpolation is first used to resample the feature maps F_i at different scales (i = 1, 2, 3, 4) to the size of input images, which are compressed into four channels via a 1×1 convolution layer, and the four channels feature maps are concatenated into a hybrid feature map $\hat{F}$. The average pooling layer with MLP is then applied to obtain a coefficient for every channel. The scale coefficient vector is denoted as γ. Next, the $\hat{F} \cdot \gamma$ is taken as input to be passed through one 3×3 convolutional layer, a ReLU activation function and a 1×1 convolutional layer, to generate the coefficient γ^*. With the aim of facilitating the training, the residual connections are used, as shown in *Figure 5*. The final output of multiscale attention structure is as follows:

$$y_{MSA} = \hat{F} \cdot \gamma \cdot \gamma^* + \hat{F} \cdot \gamma + \hat{F} \quad [5]$$

Implementation details

The experiments in the study were configured with the Windows 11 operating system (Microsoft Corp., Redmond, WA, USA) with 16 GB of memory, an NVIDIA RTX 3060 (Nvidia Corp., Santa Clara, CA, USA) with 12 GB of memory, an Intel i5-12400F chip (Intel Corp., Santa Clara, CA, USA), and a PyTorch framework based on the Python computer language. In the training process, an Adam optimizer was employed with default parameters. The training was configured for 100 epochs with a batch size of

8. The learning rate and momentum were set to 0.0003 and 0.9, respectively.

Metrics

To evaluate the performance of segmentation using the proposed network, the segmentation results on the test set were compared with the ground truth created by the experts. Five metrics were applied to quantitatively calculate the distance between the ground truth and the results segmented by the proposed network. Different metrics involved a number of evaluation criteria and are described below.

The Dice coefficient is one of the most widely used evaluation metrics for medical image segmentation and is used to calculate the overlap index between segmentation results (*SR*) and ground truth (*GT*).

$$Dice(SR,GT)=\frac{2\times|SR\cap GT|}{|SR|+|GT|} \quad [6]$$

where the Dice coefficient value is in the range of [0, 1], and if the segmentation is completely accurate, the Dice coefficient value is 1.

The volumetric overlap error (VOE) is the volume overlap error, which is opposite to the Dice coefficient. The smaller the value is, the better the segmentation performance of the network. The VOE is calculated as follows:

$$VOE=1-\frac{|SR\cap GT|}{|SR\cup GT|} \quad [7]$$

The relative volume difference (RVD) represents the difference between the segmentation results and the ground truth of volumes, which can be calculated as follows:

$$RVD(SR,GT)=\frac{|SR|-|GT|}{|GT|} \quad [8]$$

The average symmetric surface distance (ASD) indicates the average surface distance between *SR* and *GT*.

When *S*(*GT*) is the set of slice voxels of *GT*, the closest distance of any voxel v to *S*(*GT*) can be calculated as follows:

$$D(v,S(GT))=\min_{S_{GT}\in S(GT)} D(v,S(GT)) \quad [9]$$

where $D\ (v,\ S_{GT})$ is the Euclidean distance of the voxels involving the real spatial resolution of the images. The ASD can be calculated as follows:

$$ASD(GT,SR)=\frac{1}{|S(GT)|+|S(MS)|}\left(\sum_{S_{GT}\in S(GT)} D(S_{GT},S(SR))+\sum_{S_{SR}\in S(SR)} D(S_{SR},S(GT))\right) \quad [10]$$

The maximum surface distance (MSD) represents the maximum distance of the nearest points between the *SR* and *GT*, and can be calculated as follows:

$$MSD(SR,GT)=\max\left\{D\left[S_{GT},S(SR)\right],D\left[S_{SR},S(GT)\right]\right\} \quad [11]$$

Results

In this section, we present the outcomes of our study, including comparisons with other methods and ablation studies.

Comparisons with other methods

The liver and liver tumor segmentation results were evaluated for each test method using Dice, VOE, RVD, ASD, and MSD criteria. The quantitative results are summarized in *Tables 1,2*. Paired *t* tests were used to investigate whether statistical differences existed between the different models. The P values of the paired *t* tests between the Dice of proposed network and other networks are shown in the last column of *Tables 1,2*. In the liver tumor segmentation experiment, we used a 95% confidence interval. The confidence intervals for the Dice, VOE, RVD, ASD, and MSD metrics were 0.592–0.834, 0.17–0.61, −0.37 to 0.31, 0.89–4.19 mm, and 6.97–19.87 mm, respectively.

Five common methods were selected for comparison with the proposed network. The liver segmentation results are presented in *Table 1*, and the best scores for each metric are underlined. As can be seen in *Table 1* and *Figure 8*, the proposed network outperformed the other methods in terms of the Dice, RVD, ASD, and MSD scores. The Dice, VOE, RVD, ASD, and MSD scores of our method were on average 0.059, 0.05, 0.15, 0.28, and 2.32 higher than those of the UNet baseline, respectively.

The results of the liver tumor segmentation performance are shown in *Table 2* and *Figure 9*. Based on the experiment results, it is apparent that liver tumors are more difficult to segment than the liver itself. Among the five networks compared, the MA-UNet obtained the highest Dice, VOE, ASD, and MSD scores, while the proposed network achieved the best performance for the RVD metric. Although the proposed network did not yield the best performance, compared with the other four methods, it can be considered a success. The Dice, VOE, RVD, ASD,

Table 1 Results of the different methods for liver segmentation

Method	Dice	VOE	RVD	ASD (mm)	MSD (mm)	P
UNet (15)	0.912	0.13	0.17	1.29	9.13	<0.001
UNet++ (16)	0.938	0.12	0.05	1.58	8.52	<0.001
Res-UNet (47)	0.911	0.11	−0.15	1.13	9.15	<0.001
MA-UNet (48)	0.960	0.08	−0.03	1.04	7.14	<0.001
Attention-UNet (46)	0.941	0.11	−0.03	1.09	7.27	<0.001
Proposed method	0.971	0.08	0.02	1.01	6.81	–

The best results for each metric are underlined. VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance.

Table 2 Results of the different methods for tumor segmentation

Method	Dice	VOE	RVD	ASD (mm)	MSD (mm)	P
UNet (15)	0.621	0.53	−0.23	3.59	17.76	<0.001
UNet++ (16)	0.705	0.32	−0.21	2.68	12.38	<0.001
Res-UNet (47)	0.674	0.48	0.20	2.93	14.25	<0.001
MA-UNet (48)	0.729	0.21	−0.19	1.84	8.39	<0.001
Attention-UNet (46)	0.702	0.35	−0.20	2.09	10.33	<0.001
Our method	0.721	0.37	0.16	1.88	9.12	–

The best results for each metric are underlined. VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance.

and MSD scores of the proposed method were on average 0.1, 0.16, 0.07, 1.71, and 8.64 higher than those the UNet baseline, respectively.

The Dice and ASD scores of the liver tissue and tumor segmentation results were selected to compared the different methods. As shown in *Figure 10*, the overall performance of tumor segmentation was inferior to that of liver segmentation. Specifically, UNet exhibited the most marked decline in performance, with a decrease of 0.291 in Dice score and an increase of 2.3 mm in ASD. In contrast, the networks incorporating attention mechanisms performed better, indicating that the simple encoder-decoder structure struggles to handle the diversity in morphological characteristics of tumors.

Figure 11 presents a visual comparison of the segmentation results for liver tumors of different sizes between the proposed network and other methods. The first row shows the ground truth, while subsequent rows depict the segmentation results for each method. As shown in *Figure 10*, each method could generally segment the liver region, but the segmentation results for the tumor region were not satisfactory. This is consistent with the results in *Figure 9*. Additionally, in *Figures 11,12*, the small tumors in Pictures 2, 3, and 4 are barely segmented, which might have contributed to the poor segmentation results of tumors.

With the aim of evaluating the inference efficiency of the proposed network on common medical devices, we employed a general-purpose to test the network inference time. The CPU model used for bench-marking was an Intel i7-9750H operating at 2.6 GHz, which reflected the use needs of the point-of-care medical equipment. The 21 images from the LiTS2017 datasets were used to test the proposed network. The results are presented in *Table 3*. It can be clearly observed that the proposed network obtained the best performance among networks compared in this experiment. The number of parameters and inference time of the proposed network were only 5.3% and 22.1% of those of the UNet baseline. It can be also seen from *Table 3* that the inference time of the Attention-UNet with the attention gate mechanism only increased by 6.3% as compared to the UNet. This suggests that adding an attention mechanism to the basic network will not drastically increase the inference

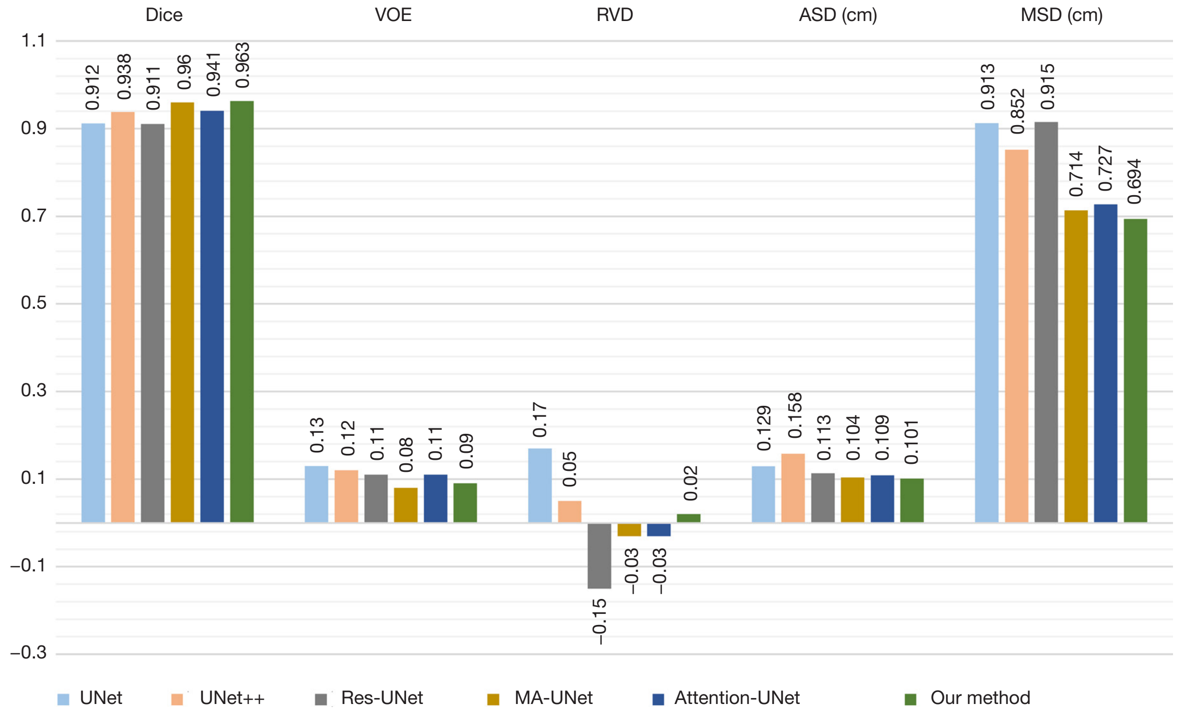

Figure 8 Quantitative comparison of the different methods for liver segmentation. VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance.

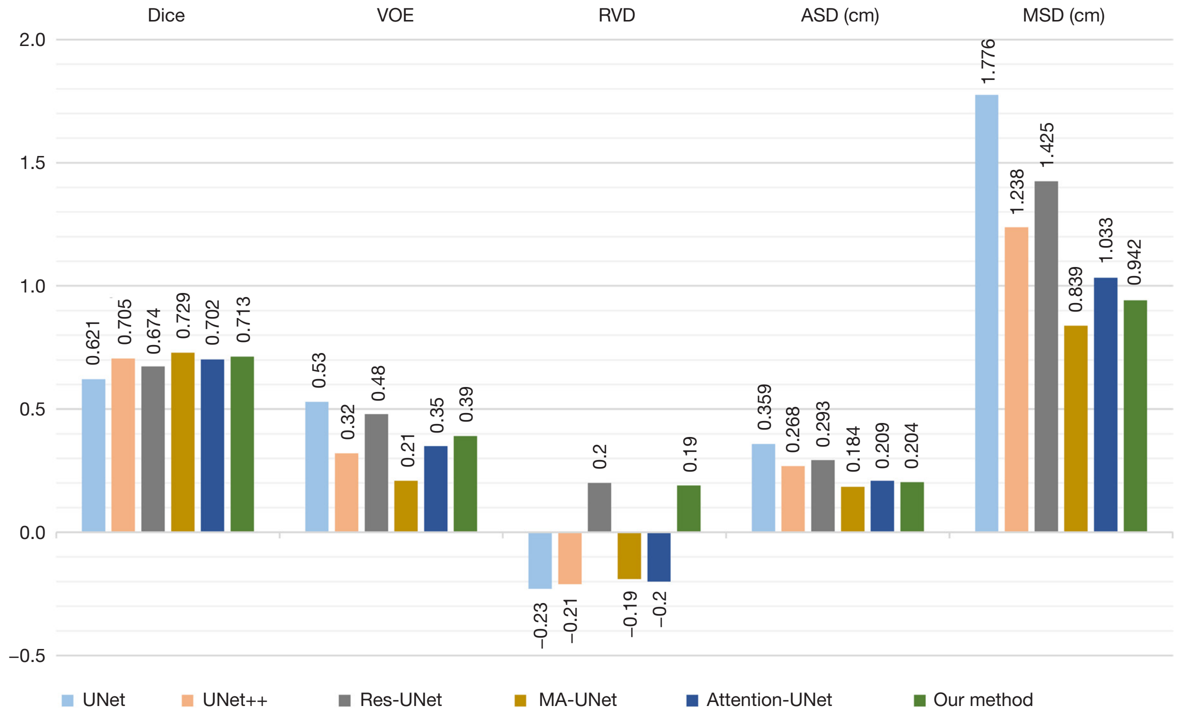

Figure 9 Quantitative comparison of different methods for tumor segmentation results. VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance.

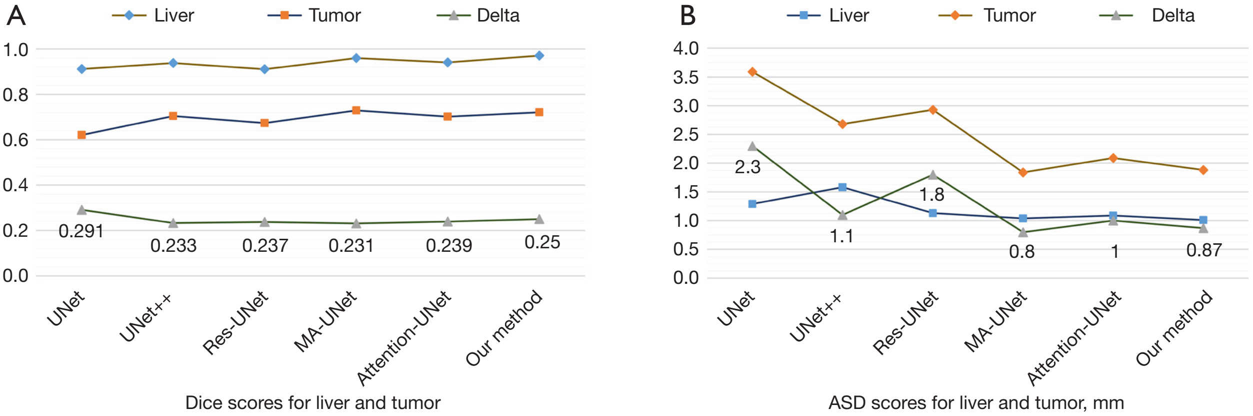

Figure 10 Comparison of the Dice and ASD scores for the liver and tumor. ASD, average symmetric surface distance.

time of the network, supporting the inclusion of attention gates into our network.

To evaluate the segmentation efficiency of networks, comparison charts of Dice score and inference time were plotted. As can clearly be seen in *Figure 13*, the MA-UNet and the proposed method garnered the best performance in terms Dice score. Moreover, the proposed network outperformed all the other networks in terms of inference time. This indicates that the proposed method is the most efficient network among those tested.

In the liver tumor segmentation experiments, as shown in *Table 2*, the proposed method performed slightly worse than did the MA-UNet in terms of the Dice, VOE, ASD, and MSD metrics. However, as demonstrated in *Figure 14*, a large area of adhesion between two tumors in the liver tumor segmentation results of MA UNet was present, which did not appear in the results of the proposed method. To further analyze this phenomenon, 3D visualization of the segmentation results from MA-UNet and the proposed method was conducted. As illustrated in *Figure 15*, MA UNet was prone to misidentifying normal tissue regions between two nearby tumors as tumor regions, deviating from the ground truth. Based on the above analysis, the morphology of liver tumors segmented by our method better aligns with reality as compared to those delineated by MA-UNet.

Ablation study for spatial shift

To verify the impact of the spatial shift operation on segmentation performance in the modified tokenized MLP block, an ablation experiment was designed with Net_block_1, Net_block_2, Net_block_3, and Net_block_4, as detailed in *Table 4*. The modified tokenized MLP block in Net_block_1 does not have a spatial shift operation in the shifted MLP width block or the shifted MLP height block. The modified tokenized MLP block of Net_block_2 has a spatial shift operation only in the shifted MLP width block. The modified tokenized MLP block in Net_block_3 has a spatial shift operation only in the shifted MLP height block. The modified tokenized MLP block in Net_block_4 has a spatial shift operation in the shifted MLP width block and the shifted MLP height block.

Table 5 presents the results of the ablation study for the modified tokenized MLP block in the liver tumor segmentation task. With Net_block_1 serving as a baseline, the performance of Net_block_2, Net_block_3, and Net_block_4 showed improvement across all metrics. From the analysis of the Dice, VOE, RVD, ASD, and MSD metrics, it could be concluded that adding spatial shift operation into the shifted MLP width block and the shifted MLP height block, respectively, can slightly improve the segmentation performance. However, adding spatial shift operation to both blocks simultaneously can substantially improve performance. From the analysis of the parameter and inference time, Net_block_4 only increased the number of parameters by 2.17% and inference time by 19 ms as compared to Net_block_1. This indicates that spatial shift operation can improve segmentation performance with minimal computational complexity.

Ablation study for attention modules

To verify the impact of various attention modules on segmentation performance, an ablation experiment was with Net_attenion_1, Net_attenion_2, Net_attenion_3, and Net_attenion_4, as outlined in *Table 6*. The Net_attenion_1

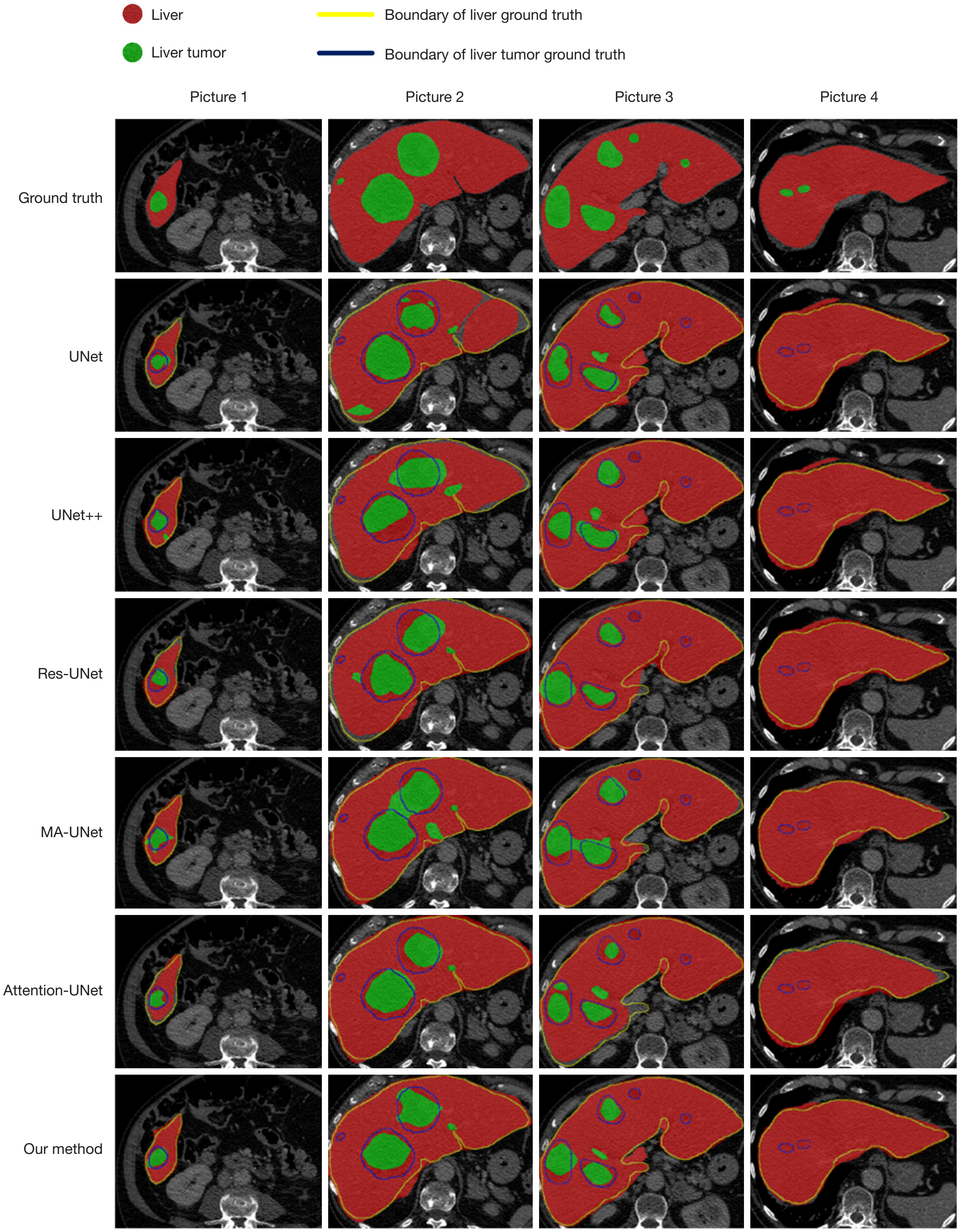

Figure 11 Comparison of the visualized results for liver and liver tumor segmentation. The red and green regions represent the liver segmentation and liver tumor segmentation results respectively, while the closed yellow and blue lines represent the boundaries of the liver and tumor ground truth, respectively.

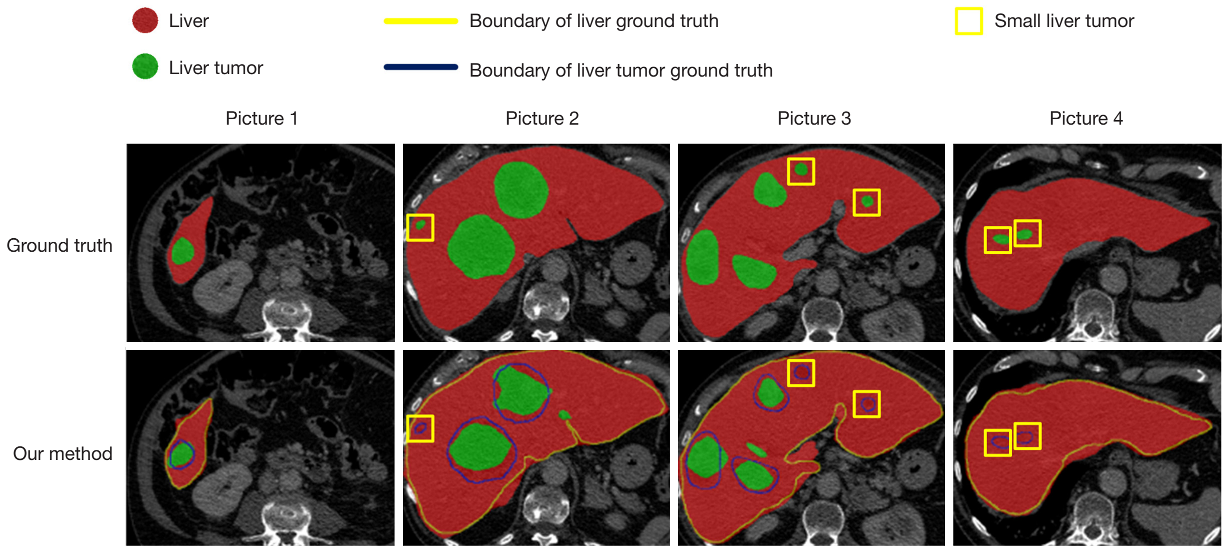

Figure 12 Visualized results of small liver tumor segmentation. The red and green regions represent the liver segmentation and tumor segmentation results, respectively, while the closed yellow and blue lines represent the boundaries of the liver and tumor ground truth, respectively. The yellow box indicates the region of the small liver tumors.

Table 3 Number of parameters and inference time of the different methods

Method	Parameters, n	Inference time (ms)
UNet	30,371,378	253
UNet++	9,163,362	173
Res-UNet	62,749,565	352
MA-UNet	34,946,839	285
Attention-UNet	34,878,638	269
Our method	1,597,803	56

The best results for each metric are underlined.

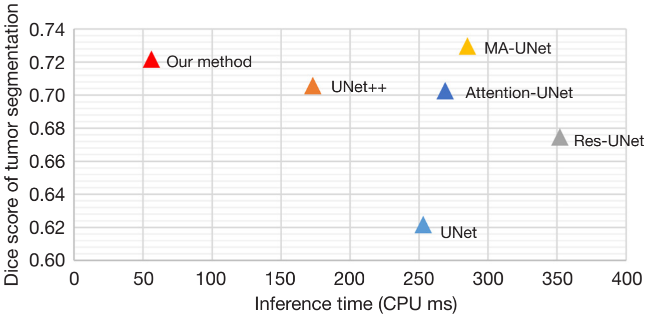

Figure 13 Comparison of liver tumor segmentation efficiency.

is the proposed network without the attention gate module and the multiscale attention module. The Net_attenion_2 is the proposed network with only the attention gate module. The Net_attenion_3 is the proposed network with only the multiscale attention module. The Net_attenion_4 is the proposed network with the attention gate module and the multiscale attention module. Notably, the modified tokenized MLP block in this iteration integrates the spatial shift operation in both the shifted MLP width block and the shifted MLP height block.

The results of the ablation study for attention modules are presented in *Table 7* for the liver tumor segmentation task. With the performance of the Net_attention_1 serving as a baseline, the Dice scores of the Net_attention_2, Net_attention_3, and Net_attention_4 were on average 0.067, 0.03, and 0.099 higher than the baseline, respectively. The MSD scores of Net_attention_2, Net_attention_3, and Net_attention_4 were on average 2.59, 3.44, and 10.19 lower than the baseline, respectively. In terms of the VOE, RVD, and ASD metrics, the networks embedded with attention modules demonstrated better performance as compared to the baseline. It is worth noting that the segmentation performance of Net_attention_2 was superior to that of Net_attention_3. From the analysis of the Dice, VOE, RVD, ASD, and MSD metrics, it is apparent that both the multiscale attention module and the attention

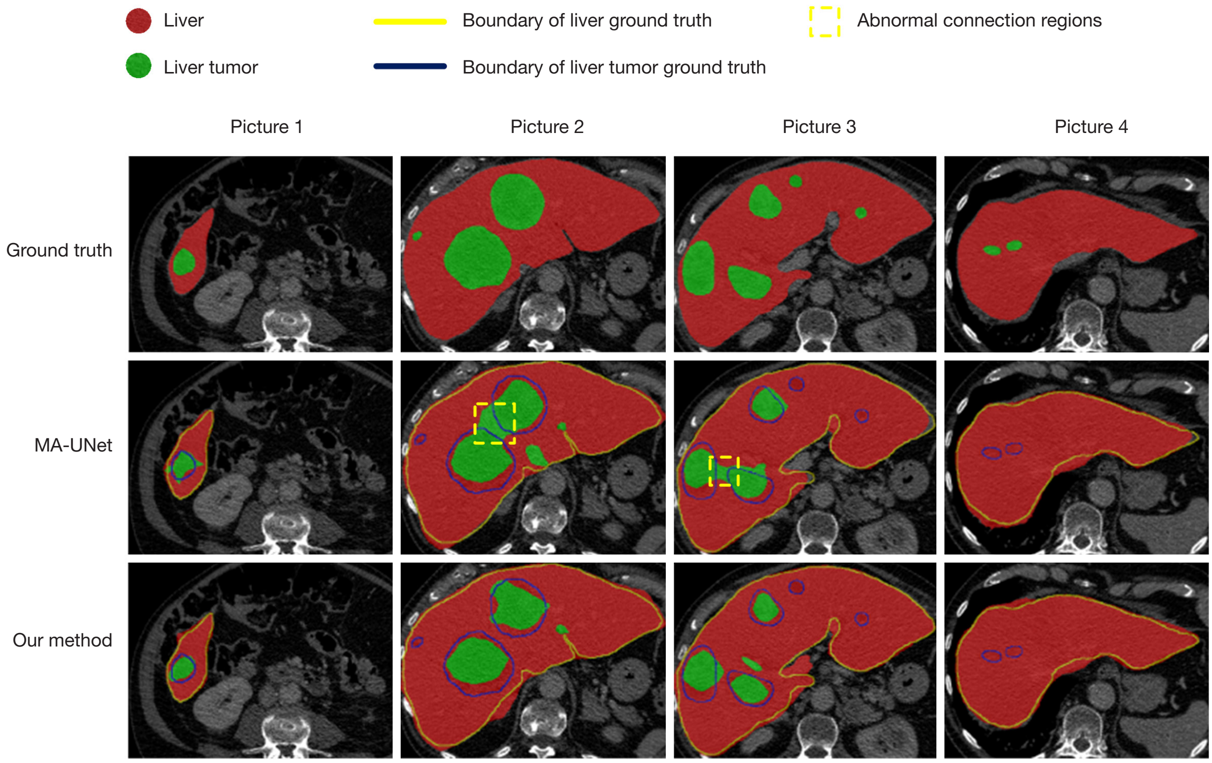
Liver
Liver tumor
Boundary of liver ground truth
Boundary of liver tumor ground truth
Abnormal connection regions
Picture 1
Picture 2
Picture 3
Picture 4
Ground truth
MA-UNet
Our method

Figure 14 Visual comparison of liver tumor segmentation results between MA-UNet and the proposed method. The red and green regions represent the liver segmentation and tumor segmentation results respectively, while the closed yellow and blue lines represent the boundaries of the liver and tumor ground truth, respectively. The yellow dashed box indicates the abnormal connection area.

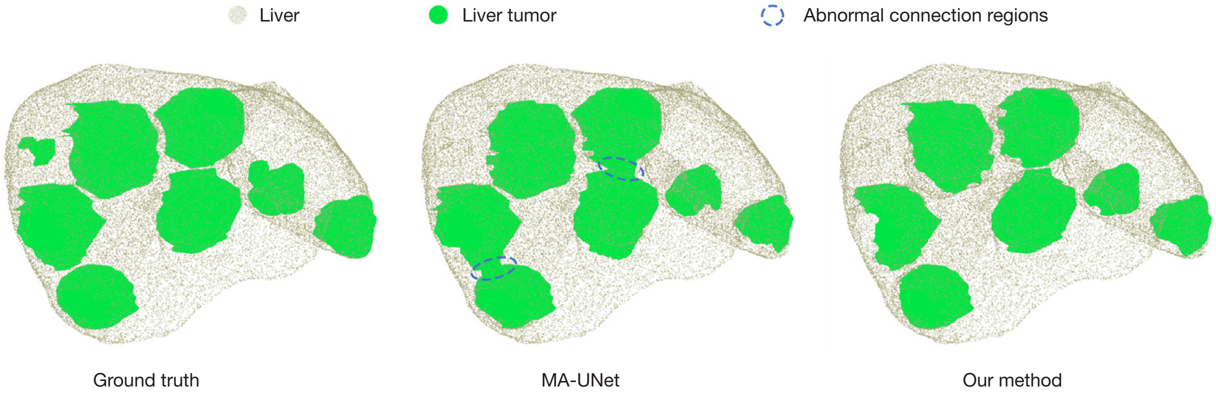

Figure 15 Three-dimensional visual comparison of liver tumor segmentation results between the MA-UNet and our method. The dark-brown point cloud represents liver tissue. The green region indicates liver tumors. The blue dashed line highlights the abnormal connection area.

gates module can promote the segmentation performance of the network. In addition, the multiscale attention module and the attention gates module did not considerably increase the number of parameters or inference time of the segmentation network. In particular, the multiscale attention module only increases the number of parameters by 2.3% and the inference time by 10 ms. In summary, both modules can improve segmentation performance while maintaining the efficiency of the network.

Discussion

In the experiment, various methods were compared in terms of their ability to automatically segment the liver and liver tumor. The results showed that the proposed network achieved the best performance in liver segmentation but suboptimal performance in tumor segmentation as compared to the other methods. Moreover, the segmentation performance for the liver was considerably superior to that for the tumor, and this phenomenon occurred for all the other networks. This could be attributed to the relatively fixed shape of the liver, with clear boundaries and contours, resulting in consistent features that are easier for the network to learn. Conversely, the variable morphologies of tumors present segmentation challenges. To further investigate the underlying reasons for this phenomenon, the segmentation results of the liver and tumors were visualized. The visualized results indicated that the network incorporating the attention gate mechanism yields considerably improved segmentation of tumor boundaries as compared to the other networks. This can be attributed to the effective learning of low-level features through the attention gate module within skip connections. Furthermore, compared to Attention-UNet (46), our network demonstrated superior capability in segmenting tumors of various sizes, indicating that assigning different weights to layers is a viable approach for learning the relevance between objects and different scale feature maps. Furthermore, although the proposed method performed slightly worse than did the MA-UNet (48) in some quantitative metrics of liver tumor segmentation, MA-UNet exhibited distortion in the 3D visualization results of liver tumors. This phenomenon may be attributed to MA-UNet's reliance on excessive multiscale feature learning, which leads to overfitting of the model and compromises the shape accuracy of liver tumor segmentation.

Table 4 Settings of ablation study for spatial shift in the modified tokenized MLP block

Method	Shifted MLP width block with spatial shift	Shifted MLP height block with spatial shift
Net_block_1		
Net_block_2	✓	
Net_block_3		✓
Net_block_4	✓	✓

MLP, multilayer perceptron.

Compared to the traditional UNet architecture (15,16,46-48), our network reduces the number of convolutions in the shallow stages and employs a modified tokenized MLP block in the deeper stages to decrease the network parameters. To evaluate the segmentation efficiency of the network, the number of parameters was computed and the inference time of the network was tested on a CPU. According to the results, our network achieved good segmentation performance with an extremely low inference time. This indicates that the proposed network architecture, with a small number of parameters, not only effectively handles high-resolution, low-level features such as tumor boundaries, shapes, and sizes but also learns semantic information from low-resolution high-level features.

In the ablation experiment, we initially investigated the impact of the spatial shift operation at different positions within the modified tokenized MLP block. The experimental results indicated that incorporating the spatial shift operation into both the shifted MLP width block and the shifted MLP height block can improve the segmentation performance of the model at the expense of only a small increase in the number of network parameters. It is

Table 5 Results of the ablation study for spatial shift in the modified tokenized MLP block

Method	Dice	VOE	RVD	ASD (mm)	MSD (mm)	Parameters, n	Inference time (ms)
Net_block_1	0.713	0.39	0.19	2.04	9.42	1,563,883	37
Net_block_2	0.716	0.37	0.17	1.98	9.28	1,585,543	43
Net_block_3	0.715	0.38	0.17	2.01	9.25	1,581,433	42
Net_block_4	0.721	0.37	0.16	1.88	9.12	1,597,803	56

VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance. The best results for each metric are underlined.

worth noting that despite a lack of spatial shift operation parameters, there was still a slight increase in the parameters when the processed feature layer is connected with the MLP. We then investigated the effect of the attention gates module and the multiscale attention module on the network segmentation performance. The experimental results in *Table* 7 demonstrate that both modules effectively enhance segmentation performance. This indicates that assigning different weights to feature maps can improve network performance at a relatively small computational cost. We also observed that the network with only the attention gate module generally outperformed the network with only the multiscale attention module. This can be explained by the attention gate module's highlighting of crucial low-level features and high-level semantic information between the encoder and decoder, enabling the acquisition of boundary information for segmentation targets. Meanwhile, in terms of MSD metrics, the performance of the network with only the multiscale attention module was superior. This indicates that the multiscale attention module renders the network more sensitive to the size information of the segmentation targets.

Furthermore, our network achieved efficient segmentation of liver tumors and can be applied not only to real-time intraoperative segmentation tasks but also to other medical image segmentation tasks, such as skin lesion segmentation involving multiscale targets.

Table 6 Setting of the attention modules in the ablation study

Method	Attention gate module	Multiscale attention module
Net_attenion_1		
Net_attenion_2	✓	
Net_attenion_3		✓
Net_attenion_4	✓	✓

However, some limitations should be acknowledged regarding this study. First, to enhance the learning efficiency of features between the encoder and decoder, we embedded attention gate mechanisms in the skip connections to suppress irrelevant regions while emphasizing salient features beneficial for segmentation tasks. Despite this, the main features transmitted through skip connections differ between shallow and deep networks: shallow skip connections primarily convey high-resolution low-level features, while deep skip connections transmit low-resolution high-level features. These differences have varying impacts on segmentation tasks. Therefore, employing the same attention gate mechanism across both shallow and deep networks may not fully leverage the unique characteristics of these features. In future work, designing an attention mechanism that specifically caters to the characteristics of the features in shallow and deep networks could be considered to improve liver tumor segmentation performance.

Second, to capture long-range dependencies, we introduced the shifted MLP operation during the simplification of the tokenized MLP network. However, the shifting was performed only in the height and width directions. In future work, applying the shifting operation in multiple directions could be considered to capture more comprehensive long-range dependencies.

Third, CT images exhibit significant sequential characteristics, with clear correlations of liver tumor information between different slices. In future research, designing networks specifically addressing the sequential nature of CT data could be considered to further improve the segmentation efficiency.

Table 7 Results of ablation study for the attention modules

Method	Dice	VOE	RVD	ASD (mm)	MSD (mm)	Parameters, n	Inference time (ms)
Net_attention_1	0.622	0.42	0.21	4.11	19.31	1,491,938	29
Net_attention_2	0.689	0.38	0.17	2.11	16.72	1,560,830	43
Net_attention_3	0.652	0.40	0.18	2.92	15.87	1,527,413	39
Net_attention_4	0.721	0.37	0.16	1.88	9.12	1,597,803	56

The best results for each metric are underlined. VOE, volumetric overlap error; RVD, relative volume difference; ASD, average symmetric surface distance; MSD, maximum surface distance.

Conclusions

This paper proposes a network based on modified tokenized MLP and attention mechanism for the efficient CT image-based segmentation of liver tumors. The proposed network uses a spatial shift operation to promote information exchange between different tokens in tokenized MLP blocks; additionally, it leverages attention gates and a multiscale attention module to enhance the feature representation of liver tumors with minimal computational complexity. The experimental indicated that the proposed network has the shortest inference time while maintaining outstanding performance in liver tumor segmentation as compared to the traditional networks. Notably, the number of parameters and inference time of our network were only 5.3% and 22.1% those of UNet, respectively. Moreover, our model could be extended to segmentation tasks of other pathological tissues and applied in real-time intraoperative segmentation scenarios. In future work, we will investigate methods to mitigate the issue of excessive loss of feature information for small targets during the downsampling process and endeavor to address the challenge of mining correlation information between sequential images.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD + AI reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-24-2132/rc

Funding: This study was supported by the National Natural Science Foundation of China (No. 52375486), the Fundamental Research Funds for the Central Universities (No. 2232024G-14), and the Donghua University 2024 Cultivation Project of Discipline Innovation (No. xkcx-202403).

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-24-2132/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The data set used in this study was acquired from a public data set, and thus the requirement for informed consent and ethical approval in this analysis was waived.

References

1. Donne R, Lujambio A. The liver cancer immune microenvironment: Therapeutic implications for hepatocellular carcinoma. Hepatology 2023;77:1773-96.
2. Sethi G, Rath P, Chauhan A, Ranjan A, Choudhary R, Ramniwas S, Sak K, Aggarwal D, Rani I, Tuli HS. Apoptotic Mechanisms of Quercetin in Liver Cancer: Recent Trends and Advancements. Pharmaceutics 2023;15:712.
3. Velichko YS, Gennaro N, Karri M, Antalek M, Bagci U. A Comprehensive Review of Deep Learning Approaches for Magnetic Resonance Imaging Liver Tumor Analysis. Adv Clin Radiol 2023;5:1-15.
4. Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G, et al. The Liver Tumor Segmentation Benchmark (LiTS). Med Image Anal 2023;84:102680.
5. Valanarasu JMJ, Patel VM. UNeXt: MLP-Based Rapid Medical Image Segmentation Network. In: Medical Image Computing and Computer Assisted Intervention (MICCAI) 2022;13435:23-33.
6. Karpiński R, Krakowski P, Jonak J, Machrowska A, Maciejewski M. Comparison of Selected Classification Methods Based on Machine Learning As a Diagnostic Tool for Knee Joint Cartilage Damage Based on Generated Vibroacoustic Processes. Appl Comput Sci 2023;19:136-50.
7. Machrowska A, Karpiński R, Maciejewski M, Jonak J, Krakowski P, Syta A. Application of Recurrence Quantification Analysis in the Detection of Osteoarthritis

of the Knee with the Use of Vibroarthrography. Adv Sci Technol Res J 2024;18:19-31.

8. Seo H, Huang C, Bassenne M, Xiao R, Xing L. Modified U-Net (mU-Net) With Incorporation of Object-Dependent High Level Features for Improved Liver and Liver-Tumor Segmentation in CT Images. IEEE Trans Med Imaging 2020;39:1316-25.
9. Mahalaxmi G, Tirupal T, Shanawaz S. Liver Cancer Detection Using Various Image Segmentation Approaches: A Review. IUP J Telecommun 2021;13:48-61.
10. Ansari MY, Abdalla A, Ansari MY, Ansari MI, Malluhi B, Mohanty S, Mishra S, Singh SS, Abinahed J, Al-Ansari A, Balakrishnan S, Dakua SP. Practical utility of liver segmentation methods in clinical surgeries and interventions. BMC Med Imaging 2022;22:97.
11. Abdel-Basset M, Chang V, Mohamed R. A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems. Neural Comput Appl 2021;33:10685-718.
12. Biratu ES, Schwenker F, Debelee TG, Kebede SR, Negera WG, Molla HT. Enhanced Region Growing for Brain Tumor MR Image Segmentation. J Imaging 2021;7:22.
13. Mittal H, Pandey AC, Saraswat M, Kumar S, Pal R, Modwel G. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. Multimed Tools Appl 2022;81:35001-26.
14. Wang R, Lei T, Cui R, Zhang B, Meng H, Nandi AK. Medical image segmentation using deep learning: A survey. IET Image Process 2022;16:1243-67.
15. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science 2015;9351:234-41.
16. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018;11045:3-11.
17. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen YW, Wu J. UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020:1055-9.
18. McHugh H, Talou GM, Wang A. 2D Dense-UNet: A Clinically Valid Approach to Automated Glioma Segmentation. In: Crimi A, Bakas S. editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes). Cham: Springer, 2021;12659:69-80.
19. Yu G, Dong J, Wang Y, Zhou X. RUC-Net: A Residual-Unet-Based Convolutional Neural Network for Pixel-Level Pavement Crack Segmentation. Sensors (Basel) 2022.
20. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, Yang Z, Zhang Y, Tao D. A Survey on Vision Transformer. IEEE Trans Pattern Anal Mach Intell 2023;45:87-110.
21. Xu H, Xu Q, Cong F, Kang J, Han C, Liu Z, Madabhushi A, Lu C. Vision Transformers for Computational Histopathology. IEEE Rev Biomed Eng 2024;17:63-79.
22. Li Z, Li Y, Li Q, Wang P, Guo D, Lu L, Jin D, Zhang Y, Hong Q. LViT: Language Meets Vision Transformer in Medical Image Segmentation. IEEE Trans Med Imaging 2024;43:96-107.
23. He A, Wang K, Li T, Du C, Xia S, Fu H. H2Former: An Efficient Hierarchical Hybrid Transformer for Medical Image Segmentation. IEEE Trans Med Imaging 2023;42:2763-75.
24. He Q, Duan Y, Yang Z, Wang Y, Yang L, Bai L, Zhao L. Context-aware augmentation for liver lesion segmentation: shape uniformity, expansion limit and fusion strategy. Quant Imaging Med Surg 2023;13:5043-57.
25. Wang F, Cheng XL, Luo NB, Su DK. Attention-guided context asymmetric fusion networks for the liver tumor segmentation of computed tomography images. Quant Imaging Med Surg 2024;14:4825-39.
26. Liu T, Liu J, Ma Y, He J, Han J, Ding X, Chen CT. Spatial feature fusion convolutional network for liver and liver tumor segmentation from CT images. Med Phys 2021;48:264-72.
27. Luan S, Xue X, Ding Y, Wei W, Zhu B. Adaptive Attention Convolutional Neural Network for Liver Tumor Segmentation. Front Oncol 2021;11:680807.
28. Meng L, Zhang Q, Bu S. Two stage liver and tumor segmentation algorithm based on convolutional neural network. Diagnostics 2021;11:1806.
29. Cheng D, Zhou Z, Zhang J. EG-UNETR: An edge-guided liver tumor segmentation network based on cross-level interactive transformer. Biomed Signal Process Control 2024;97:106739.
30. Li R, Xu L, Xie K, Song J, Ma X, Chang L, Yan Q. DHT-Net: Dynamic Hierarchical Transformer Network for Liver and Tumor Segmentation. IEEE J Biomed Health Inform 2023;27:3443-54.
31. Di S, Zhao YQ, Liao M, Zhang F, Li X. TD-Net: A Hybrid End-to-End Network for Automatic Liver Tumor Segmentation From CT Images. IEEE J Biomed Health Inform 2023;27:1163-72.
32. Tolstikhin I, Houlsby N, Kolesnikov A, Beyer L, Zhai X,

Unterthiner T, Yung J, Steiner A, Keysers D, Uszkoreit J, Lucic M, Dosovitskiy A. MLP-Mixer: An all-MLP Architecture for Vision. Adv Neural Inf Process Syst 2021;29:24261-72.

33. Cheng Z, Qu A, He X. Contour-aware semantic segmentation network with spatial attention mechanism for medical image. Vis Comput 2022;38:749-62.
34. Guo C, Szemenyei M, Yi Y, Wang W, Chen B, Fan C. SA-UNET: Spatial attention U-net for retinal vessel segmentation. In: 25th International Conference on Pattern Recognition (ICPR) 2020;1236-42.
35. Huang G, Zhu J, Li J, Wang Z, Cheng L, Liu L, Li H, Zhou J. Channel-Attention U-Net: Channel Attention Mechanism for Semantic Segmentation of Esophagus and Esophageal Cancer. IEEE Access 2020;8:122798-810.
36. Guo C, Szemenyei M, Hu Y, Wang W, Zhou W, Yi Y. Channel attention residual u-net for retinal vessel segmentation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021;2021-June:1185-9.
37. Gu R, Wang G, Song T, Huang R, Aertsen M, Deprest J, Ourselin S, Vercauteren T, Zhang S. CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation. IEEE Trans Med Imaging 2021;40:699-711.
38. Guo MH, Xu TX, Liu JJ, Liu ZN, Jiang PT, Mu TJ, Zhang SH, Martin RR, Cheng MM, Hu SM. Attention mechanisms in computer vision: A survey. Comput Vis Media 2022;8:331-68.
39. Chen B, Liu Y, Zhang Z, Lu G, Kong AWK. TransAttUnet: Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation. IEEE Trans Instrum Meas 2024;8:55-68.
40. Zhou T, Canu S, Ruan S. Automatic COVID-19 CT segmentation using U-Net integrated spatial and channel attention mechanism. Int J Imaging Syst Technol 2021;31:16-27.
41. Niyas S, Pawan SJ, Anand Kumar M, Rajan J. Medical image segmentation with 3D convolutional neural networks: A survey. Neurocomputing 2022;493:397-413.
42. Liu X, Song L, Liu S, Zhang Y. A review of deep-learning-based medical image segmentation methods. Sustain 2021;13:1224.
43. Han L, Chen Y, Li J, Zhong B, Lei Y, Sun M. Liver segmentation with 2.5D perpendicular UNets. Comput Electr Eng 2021;91:107118.
44. Zhang C, Hua Q, Chu Y, Wang P. Liver tumor segmentation using 2.5D UV-Net with multi-scale convolution. Comput Biol Med 2021;133:104424.
45. Yu T, Li X, Cai Y, Sun M, Li P. S2-MLP: Spatial-Shift MLP Architecture for Vision. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2022;3615-24.
46. Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B, Rueckert D. Attention gated networks: Learning to leverage salient regions in medical images. Med Image Anal 2019;53:197-207.
47. Rahman H, Bukht TFN, Imran A, Tariq J, Tu S, Alzahrani A. A Deep Learning Approach for Liver and Tumor Segmentation in CT Images Using ResUNet. Bioengineering (Basel) 2022;9:368.
48. Fan T, Wang G, Li Y, Wang H. Ma-net: A multi-scale attention network for liver and tumor segmentation. IEEE Access 2020;8:179656-65.

Cite this article as: Yang B, Zhang J, Lyu Y, Zhang J. Automatic computed tomography image segmentation method for liver tumor based on a modified tokenized multilayer perceptron and attention mechanism. Quant Imaging Med Surg 2025;15(3):2385-2404. doi: 10.21037/qims-24-2132