IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 30, 2020, accepted September 16, 2020, date of publication September 21, 2020, date of current version October 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3025372

MA-Net: A Multi-Scale Attention Network for Liver and Tumor Segmentation

TONGLE FAN, GUANGLEI WANG, YAN LI, AND HONGRUI WANG
College of Electronic and Information Engineering, Hebei University, Hebei 071002, China

Corresponding author: Guanglei Wang (513197133@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61473112, in part by the Hebei Provincial Natural Science Fund Key Project under Grant F2017201222, in part by the Education Department Science and Technology Research Project under Grant QN2015135, and in part by the Innovation and Entrepreneurship Program of Hebei University under Grant hbu2020ss065.

ABSTRACT Automatic assessing the location and extent of liver and liver tumor is critical for radiologists, diagnosis and the clinical process. In recent years, a large number of variants of U-Net based on Multi-scale feature fusion are proposed to improve the segmentation performance for medical image segmentation. Unlike the previous works which extract the context information of medical image via applying the multi-scale feature fusion, we propose a novel network named Multi-scale Attention Net (MA-Net) by introducing self-attention mechanism into our method to adaptively integrate local features with their global dependencies. The MA-Net can capture rich contextual dependencies based on the attention mechanism. We design two blocks: Position-wise Attention Block (PAB) and Multi-scale Fusion Attention Block (MFAB). The PAB is used to model the feature interdependencies in spatial dimensions, which capture the spatial dependencies between pixels in a global view. In addition, the MFAB is to capture the channel dependencies between any feature map by multi-scale semantic feature fusion. We evaluate our method on the dataset of MICCAI 2017 LiTS Challenge. The proposed method achieves better performance than other state-of-the-art methods. The Dice values of liver and tumors segmentation are 0.960 ± 0.03 and 0.749 ± 0.08 respectively.

INDEX TERMS CT, liver tumor segmentation, deep learning, attention mechanism, context information.

I. INTRODUCTION

Liver cancer has become one of the most common diseases for human and causes massive deaths every year [1], [2]. The liver and liver lesions are segmented manually by radiologists, which is time-consuming and depends on the expertise of the radiologists for segmentation accuracy. Therefore, automatic liver and tumors segmentation methods become critical in the clinical practice. In the past few years, Convolutional Neural Network (CNN) had achieved great success in the image segmentation field. Numerous methods based on Fully Convolutional Networks (FCN) [3] have been proposed to segment images accurately. Compared with the natural image segmentation, medical image segmentation is a huge challenging task because of the low intensity contrast between the organs and the various size, shape and location of lesion area within one patient. Moreover, some tumors have fuzzy boundaries which bring extremely complicated tasks for accuracy detection and segmentation. To tackle the above difficulties, many methods based on Deep Learning have been proposed in the medical image segmentation field. The U-Net [4] is one of the most popular network architectures which based on encoder-decoder network in the medical image segmentation field. To improve segmentation performance, it employs skip connections to exploit multi-scale information features. Moreover, many works with the latest skip connections have been proposed to improve network architecture such as residual connections [5] and densely connections [6] and U-Net++ [7]. Although the variances of skip connections proposed help to capture rich different-levels semantic features, it cannot describe spatial and channel-wise relationships between pixels of image, which are essential for medical image segmentation.

In addition to designing skip connections to fuse different-level semantic features, other state-of-the-art methods based on FCNs architecture have been proposed to capture Multi-scale context feature information of image via using dilated convolutions with different sampling rate [8]–[10] and

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

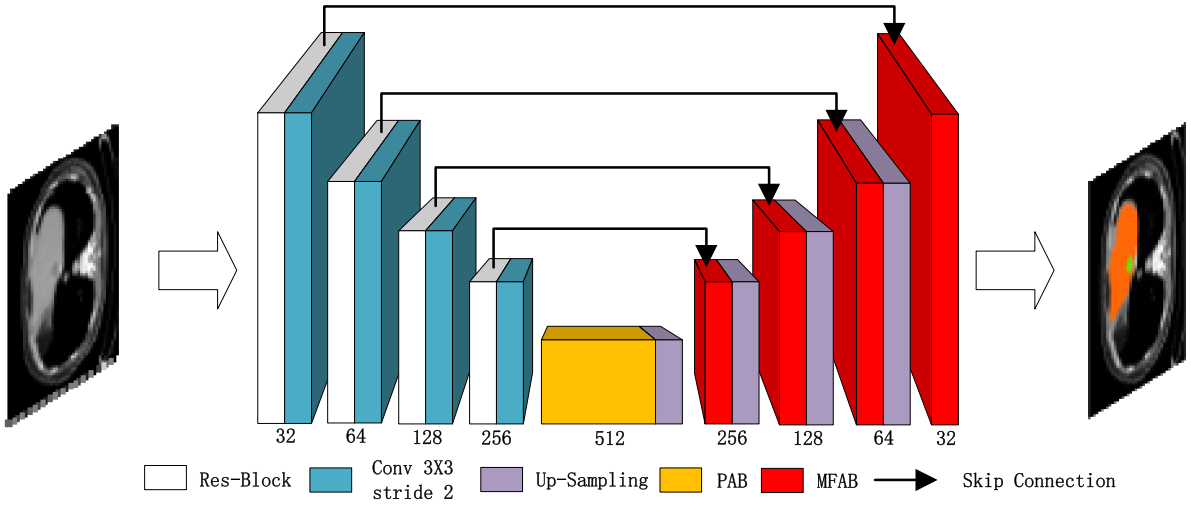


FIGURE 1. The total architecture of MA-Net.

pooling operations [11]–[14]. For example, [14] designed Residual Multi-kernel Pooling (RMP) strategy which has different-size pooling kernels to fuse multi-scale context feature information. The dilated convolutions with different sampling rate and pooling operations are used to obtain rich Multi-scale context information of images, which further improve segmentation performance. However, the dilated convolutions and pooling operations cannot leverage the spatial and channel-wise relationship between pixels in a global view. Moreover, it is easy to loss details from the feature map information by using pooling operations.

In order to address the above problems, we propose a novel network architecture named Multi-scale Attention-Net (MA-Net) for liver and tumors segmentation, which is shown in Fig1. The self-attention mechanism is used in the MA-Net. Specifically, we use two blocks based on self-attention mechanism to capture spatial and channel dependencies of feature maps. One is Position-wise Attention Block (PAB), and the other is Multi-scale Fusion Attention Block (MFAB). The PAB is used to obtain the spatial dependencies between pixels in feature maps by a self-attention mechanism manner. The MFAB is used to capture the channel dependencies between any feature maps by applying attention mechanism. Besides considering the channel dependencies of high-level feature maps, the channel dependencies of Low-level feature maps are also considered in the MFAB. The channel dependencies of high-level and low-level feature maps are fused in a sum manner, which aims to obtain rich Multi-scale semantic information of feature maps by using attention mechanism and improve network performance.

In this work, our main contributions can be summarized as follows.

(1) We propose a novel network named Multi-scale Attention-Net with the dual attention mechanism to enhance the ability of feature representation for liver and tumors segmentation.

(2) We design two Blocks with self-attention mechanism: Position-wise Attention Block (PAB) and Multi-scale Fusion Attention Block (MFAB). We use PAB and MFAB to capture attention feature maps of spatial and channel levels. The PAB is proposed to obtain the spatial dependencies between pixels in a global view, and the MFAB is to capture the channel dependencies between any feature maps by fusing high and low-level semantic features.

II. RELATED WORK

A. MULTI-SCALE INFORMATION EXTRACTION

Multi-scale information can provide rich semantic features for medical image segmentation. In the past few years, many methods [4], [12]–[17] proposed appling Multi-scale information to enhance contextual aggregation. We review several methods about Multi-scale information extraction.

The U-Net [4] have achieved great success in the medical image segmentation. The U-Net is a classical skip-net, and it uses skip connections to fuse low-level semantic features. Many methods are proposed with applying skip connections to obtain Multi-scale information of images [7], [16], [18], [19]. For example, [16] designed the cross dense connections to capture the different-level semantic features. In addition to using the common skip connections between encoder and decoder path, [18] designed a novel skip connection named high-resolution pathway, and the skip connection used dilation convolution with different rate to obtain Multi-scale information.

In addition to applying skip connections, the dilation convolution [8] with different rate and pooling operation [11] is also used to capture Multi-scale information [12], [14], [17], [20]. For instances, to gather Multi-scale information, [14] designed a pooling strategy with different-size pooling kernels for medical image segmentation. Reference [20] introduced cascaded context pyramid with dilation convolution of different dilation rate into the proposed network to

capture multi-scale semantic information. Although the skip connection, dilation convolution and pooling operation can obtain the context fusion information, they cannot describe the spatial and channel relationship between objects in a global view. It is important to gain rich semantic information on the basis of local features as well.

B. ATTENTION MECHANISM

Attention mechanism has been widely applied in many fields, such as [21]–[24]. Especially, the [24] is been seen as the first to use the attention mechanism to capture the global dependencies of inputs. Recently, the attention mechanism is popular and widely used in computer vision tasks [25]–[30]. The attention mechanism can simulate the human visual system and focus on the areas of interest. Meanwhile, attention mechanism can capture the long-range dependencies. For example, [25] designed a novel Squeeze-and-Excitation unit to obtain the channel dependencies between feature maps and adaptively recalibrates channel dependencies responses of feature maps. Reference [30] introduced the Self-Attention into Generative Adversarial Network, which models the long-range dependency for image generation tasks. Moreover, the attention mechanism is also applied in the medical image segmentation fields [9], [15], [31], [32]. Through the use of SEblock, [31] fused 2D and 3D feature maps for chronic stroke lesion segmentation. Reference [15] embedded the SEblock in the U-Net network, which obtained the channel dependencies between feature maps for prostate zonal segmentation. Reference [32] designed a novel attention module which utilized the attention mechanism for prostate segmentation. Unlike previous works, we consider the spatial and channel dependencies in our method. Moreover, we also fuse the high and low-level semantic feature in the channel-wise dependencies.

C. LIVER AND TUMOR SEGMENTATION

In recent years, many methods based on convolutional neural network have been proposed for liver and tumor segmentation. Some methods are based on 2D networks or 3D networks respectively. Reference [33] proposed a bottleneck feature supervised (BS) 2D U-Net which uses convolution kernels of different sizes to obtain multi-scale feature maps for live and tumor segmentation. Reference [34] proposed GIU-Net that combines an improved 2D U-Net neural network model with graph cutting for liver segmentation. The GIU-Net model achieved dice scores for liver segmentation is 95.05%. Reference [35] proposed 3D RA-UNet which achieved dice scores 96.1% for liver segmentation and 59.5% for tumor segmentation. Reference [6] proposed H-DenseUNet which combines 2D and 3D network for liver and tumor segmentation. The H-DenseUNet achieved dice score 96.5% for liver segmentation and 72.2% for tumor segmentation. However, the 2D network can not get more spatial information of images for liver and tumor segmentation. These methods that apply 3D network to segment liver and tumor usually take much more time to train, and these methods own much more parameters. Moreover, the results of 3D networks are sensitive to parameters initialization.

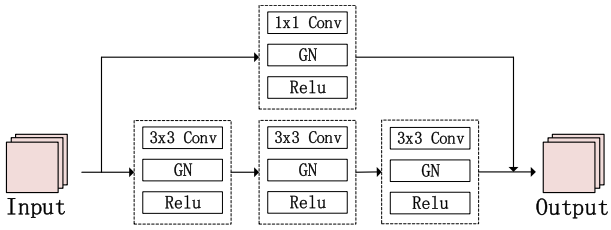


FIGURE 2. The total architecture of Res-block in our proposed method. The Res-block consists of the convolution layers, GN and residual connection.

III. MATH

In this section, we describe the proposed method in detail including Res-block, Position-wise Attention Block and Multi-scale Fusion Attention Block. We adopt the improved encoder-decoder architecture of U-Net for liver and tumors segmentation in the paper. The Res-block consists of three 3×3 Convolution blocks and residual connections to extract high-dimensional feature information. The Position-wise Attention Block is used to capture the spatial dependencies of feature maps, and the Multi-scale Fusion Attention Block is to aggregate the channel dependencies between any feature maps via fusing High and Low-level feature information.

A. RES-BLOCK

With the increasing of network layers, [5] designed a novel skip connection named residual connection to address the problem of vanishing gradient. Inspired the residual connections, we use three 3×3 Conv blocks and one residual connection to capture high-dimensional feature information of CT images in the encoder path. The 1×1 Conv is to control the number of input channels. Because the size of the experimental platform's memory is limited, the batch size usually is small in the image segmentation field. The small batch size can cause performance degradation of model. Hence, [36] proposed the group normalization to alleviate the problem. We replace Batch Normalization with group normalization in the MA-Net. We use the group normalization in the Res-block. The frame of Res-block is shown as Fig2.

B. POSITION-WISE ATTENTION BLOCK

Previous work [11], [37] have suggested that local feature information captured via using traditional convolutional network could lead to misclassification of objects. In order to capture rich contextual relationships over local feature maps, [26] designed a position attention module. Inspired the position attention module, we use PAB to capture the spatial dependencies between any two position feature maps. The PAB can model a wider range of rich spatial contextual information over local feature maps.

The frame of PAB is shown as Fig3. Given a local feature map $I \in R^{H \times W \times 256}$ as input, then we feed it into a 3×3 convolution layer to obtain $I' \in R^{H \times W \times 512}$. Then, we utilize

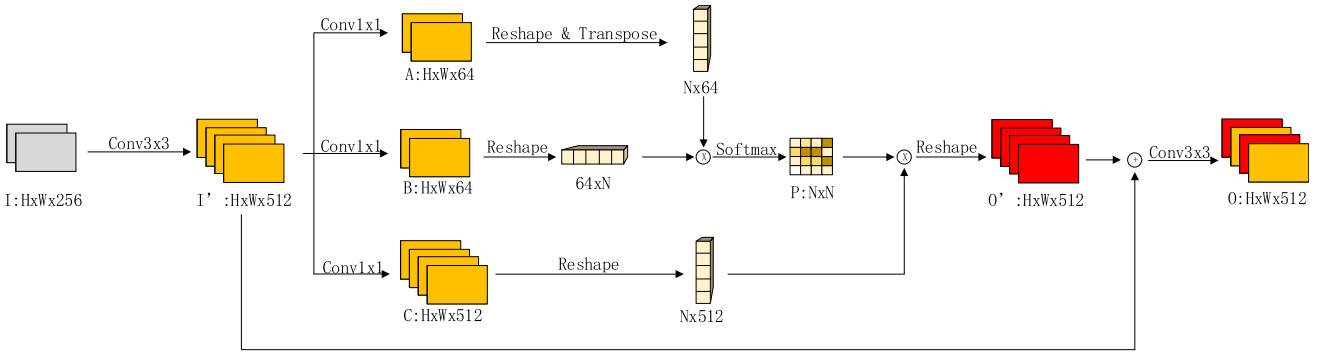


FIGURE 3. The Position-wise Attention Block (PAB). The input image is HxWx256 and output is HxWx512. The attention feature map is obtained by Softmax function.

1×1 Convolution layers to generate $A \in R^{HXWX64}$, $B \in R^{H\times W\times 64}$ and $C \in R^{H\times W\times 512}$ respectively. We reshape A and B to $R^{N\times 64}$ and R^{64XN} respectively, and then the matrix multiplication is performed between $A \in R^{N\times 64}$ and $B \in^{64\times N}$. Where N is the number of pixels. After that we use softmax function to obtain the spatial attention feature map $P \in R^{NXN}$ (see Equation(1)). Where P_{ji} denotes the i^{th} position's impact on j^{th} position in the feature map.

$$p_{ji} = \frac{\exp(A_i B_j)}{\sum_{i=1}^{N} \exp(A_i B_j)} \tag{1}$$

Meanwhile, we reshape $C \in R^{H\times W\times 512}$ to $C \in R^{N\times 512}$. We perform a matrix multiplication between the spatial attention map and $C \in R^{N\times 512}$, and reshape the result to $O' \in R^{H\times W\times 512}$. After that, we use an element-wise sum operation between I' and O'. Finally, we obtain the final output $O \in R^{H\times W\times 512}$ by a 3×3 convolution layer. The final output O is as following:

$$O_i = \alpha \sum_{i=1}^{N} (P_{ji} C_i) + I'_j \tag{2}$$

where α is initially set to 0 and gradually learns to assign more weight in the training phase. The final output O at each position is a weighted sum of the feature maps across all positions and original feature maps. Therefore, the final output O has a global contextual view and selectively aggregates rich contextual information over local feature maps according to the spatial attention map, and it considers the long-range spatial dependency between features in a global view, which improves intra-class correlation and semantic consistency.

C. MULTI-SCALE FUSION ATTENTION BLOCK

The attention mechanism in the deep learning is similar to the human visual system. It aims to select information which is important for current task from a variety of information. SENet models the channel-wise dependencies among feature channels and automatically obtains the importance of each feature channel. The purpose is to enhance the helpful feature maps and suppress the feature maps that are useless for current task. The each channel feature map of high dimensions can be seen as class-specific response. The area of liver and tumor is relatively small compared to the whole CT image. Hence, we try to imitate physicians to review CT images via introducing attention mechanism into MA-Net. By capturing the channel-wise dependencies among feature maps, model can improve the ability of feature representation. Moreover, many previous works suggest that the Multi-scale information helps to improve the segmentation accuracy.

Inspired [25], we design a novel Multi-scale Fusion Attention Block (MFAB) to extract the interdependence among feature channels via combining the High and Low-level feature maps. The MFAB is similar to the human visual system and automatically select the information that is important for liver and tumor segmentation from a variety of information. Our main idea for designing MFAB is that the MFAB learns the importance of each feature channels which come from multi-level feature maps without extra spatial dimension, and enhance the helpful feature maps and suppress feature maps that have less contribution for liver and tumor segmentation task according to the importance.

Specifically, we describe the interdependence of feature channels from Low-level and High-level feature maps. The High-level features have rich semantic information of image and the Low-level features from Skip-Connection have more edge information. The Low-level features are used to recover the details of images. The MFAB is shown as Fig4. We apply attention mechanism of channel-wise for High-level and Low-level features, respectively. The purpose is to increase the weight of important information for each feature channel in segmentation task and the useless feature information is omitted. Firstly, we obtain XH_{input} by feeding High-level feature into 1×1 and 3×3 convolution layer. The definition of is :

$$F_{Cov} : XH^*_{input} \rightarrow XH_{input} \tag{3}$$

where $XH^*_{input} \in R^{H\times W\times C'}$ and $XH_{input} \in R^{H\times W\times C}$. XH_{input} and XL_{input} have the same number of channels.We use $V = [v_1, v_2, \ldots, v_c]$ as the set of filter kernels, where v_c refers to the parameters of the c-th filter. We can calculate the

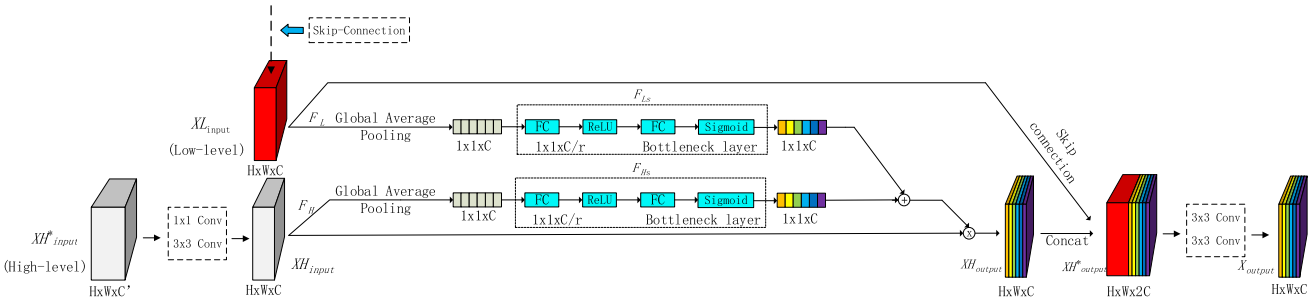


FIGURE 4. The Multi-scale Fusion Attention Block (MFAB). We use two SE-Blocks to capture Low-level and High-level feature map respectively. The final channel attention feature map is obtained via a Concat connection.

output $U = [u_1, u_2, \ldots, u_c]$:

$$u_c = v_c * X_{input} = \sum_{i=1}^{C} (v_c^i) * x^i \tag{4}$$

where $v_c = [v_c^1, v_c^2, \ldots, v_c^c]$ and $X_{input} = [x^1, x^2, \ldots, x^c]$, $X_{input} \in (XH_{input} or XL_{input})$. Here $*$ denotes convolution.

Then the global average pooling is used to compress each feature and become the number column of $1 \times 1 \times C$ and generate channel-wise statistics. Formally, the statisticS_{c1} and S_{c2} are obtained by shrinking the feature maps XH_{input} and XL_{input}. The c-th elements of S_1 and S_2 are calculated as:

$$S_{c1} = F_L(XL_{input}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{5}$$

and

$$S_{c2} = F_H(XH_{input}) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{6}$$

where H and W denotes height and width respectively, and u_c denotes feature map of each channel. Then Bottleneck layers with two Fully-Connected(FC) layers and activation function are used to limit model complexity and capture the channel-wise dependencies z_1 and z_2.

$$z_1 = F_{Ls}(S_1, P) = \delta_1(P_1 \delta_2(P_2, S_1)) \tag{7}$$

$$z_2 = F_{Hs}(S_2, P) = \delta_1(P_1 \delta_2(P_2, S_2)) \tag{8}$$

where P_1 and P_2 denote the Fully-Connected layers, $P_1 \in R^{\frac{c}{r}C}$ and $P_2 \in R^{C\frac{c}{r}}$. δ_1 and δ_2 denote sigmoid function and ReLU activation function respectively. The reduction ratio r (r=16 is best) is used to control the number of channel.

Then we use F_{add} function to combine the channel-wise output of Low-level and High-level feature.

$$z = F_{add}(\cdot) = z_1 + z_2 \tag{9}$$

The XH_{output} is obtained by rescaling T that has the activation V:

$$\tilde{X}H_{output-c} = F_{scale}(T_c, V_c) = V_c T_c \tag{10}$$

where $\tilde{X}H_{output} = [\tilde{X}H_{output-1}, \tilde{X}H_{output-2}, \ldots, \tilde{X}H_{output-c}]$ and $F_{scale}(T_c, V_c)$ is the channel-wise multiplication between the scalar V_c and the feature $T_c \in R^{H \times W}$. The multiplication can complete the re-calibration for the original feature on the channel dimension. In order to enhance the feature representation and rich semantic information, we obtain the XH^*_{output} via concatenate XL_{input} and XH_{output}. The final output X_{output} of MFAB is obtained by two 3×3 convolution layers that capture semantic information.

D. LOSS FUNCTION

The binary cross-entropy is frequently used as loss function in numerous image segmentation tasks. A loss function based on Dice had been used extensively in medical image segmentation. The Dice loss function can mitigate the imbalance problem of background and foreground pixels. However, liver and tumors are more complex and various. The target regions of tumors usually occupy smaller areas than other regions. Since the Dice loss function only pays attention to the accuracy rate in the training process, we use a weighted loss function to optimize the MA-Net. We employ the combination of cross-entropy and Dice as the final loss function in this paper. The loss function is described as:

$$L_{loss} = -\frac{1}{N} \sum_{i=1}^{N} (\alpha y_i \log p_i + \beta \frac{y_i p_i}{y_i + p_i}) \tag{11}$$

where y_i and p_i denote the ground truth and the predicted feature map, and N denotes the batch size. We use two hypeparameters ($0 < \alpha < 5$ and $0 < \beta < 5$) to control the effect of the weighted loss function. In this paper, $\alpha = 0.5$ and $\beta = 2$ can obtain the best performance.

IV. EXPERIMENTS AND RESULTS

A. PRE-PROCESSING AND IMPLEMENTATION DETAILS

We tested the proposed method on the dataset of MICCAI 2017 Liver Tumor Segmentation (LiTS) challenge [38]. The LiTS dataset consists of 131 training and 70 testing CT scans images, and the LiTS dataset provide ground truth for liver and tumors contours. For preprocessing, we truncated the pixel intensity values of all scans to the range of [−200, 250] HU to remove irrelevant tissues and enhance the contrast between liver and other tissues. For liver segmentation, the size of each CT image is 512×512 pixels and will be cropped to 256×256 to accelerate the training phase

TABLE 1. The value of Dice, VOE and RVD on the ablation analysis (standard deviation).

Method	Tumor			Liver		
	Dice	VOE	RVD	Dice	VOE	RVD
U-Net (backbone)	0.633±0.17	0.39±0.07	-0.36±0.24	0.913±0.04	0.12±0.04	-0.09±0.07
U-Net + PAB	0.719±0.11	0.21±0.04	-0.22±0.11	0.935±0.02	0.12±0.03	-0.07±0.06
U-Net + MFAB	0.738±0.06	0.19±0.05	-0.22±0.12	0.946±0.03	0.10±0.03	-0.04±0.02
MA-Net	**0.749±0.08**	**0.21±0.06**	**-0.18±0.07**	**0.960±0.03**	**0.08±0.02**	**-0.03±0.03**

and increase the region of foreground. For tumor segmentation, we use the result of liver segmentation to locate the liver. We crop out a rectangular area which contains the liver, and the rectangular with 20 pixels margin at the top, bottom, left and right. 10% of training dataset is randomly selected as the verification dataset. Moreover, we applied the data augmentation methods to avoid the overfitting problem, such as random vertical and horizontal flip and random scaling between 0.8 and 1.2 in preprocessing process. We apply the early stopping strategy as our regularization method in the training phase.

The proposed method is implemented on the platform of Pytorch. The Adam optimizer is used and the initial learning rate is set to 0.002 and decayed based on the equation $lr = lr \times (1 - iter/total\_iter)^{0.9}$. The maximal number of epoch is set to 80 and 150 for liver and tumor segmentation, respectively. All models are trained on Intel I7-9700K, Nvidia GeForce RTX 2070S with 8GB.

B. EVALUATION METRICS

To effectively evaluate the segmentation performance of the proposed method for liver and tumors. The evaluation metrics formulae as shown below:

$$Dice = \frac{2|A \cap B|}{|A| + |B|} \tag{12}$$

$$VOE = 1 - \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{13}$$

$$RVD = \frac{|A| - |B|}{|B|} \tag{14}$$

where A and B denote the predicted binary image and the ground true binary image respectively.

C. ABLATION ANALYSIS

In this section, we apply ablation analysis to evaluate the effectiveness of proposed method for liver and tumors segmentation. To evaluate the segmentation performance, we replace all original convolution layers with Res-block in the U-Net. The Res-blocks are used to extract feature information of images in the encoder path. The result of ablation analysis is shown as Table 1. The U-Net with Res-blocks is seen as backbone in the table 1. The U-Net with PAB and MFAB both improve the segmentation performance for liver and tumor segmentation. The proposed method achieves better segmentation results for liver and tumors segmentation. It proves that the PAB and MFAB are beneficial to improve the segmentation performance. With the increasing number of network depth, the segmentation accuracy will also increase. Hence, to prove that MA-Net does not rely on the number of parameters to improve the segmentation accuracy, we increase the number of parameters of the traditional U-Net. The comparative result is shown as Fig5. It shows that the MA-Net still achieves better segmentation accuracy compared to the U-Net+wide. The MA-Net achieves 0.749 and 0.960 (Dice) on the tumor and liver segmentation results by the measurement of dice respectively, compared to the U-Net+wide which achieves 0.625 and 0.913 (Dice) respectively.

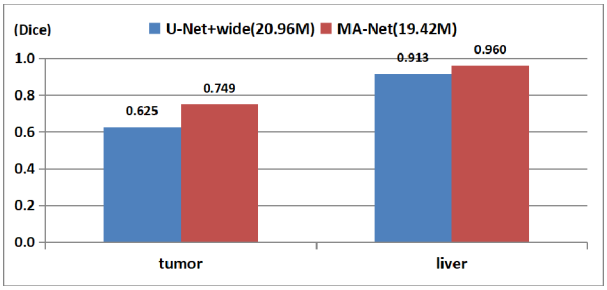


FIGURE 5. The comparative result of MA-Net and U-Net+wide.

D. COMPARISON WITH OTHER STATE-OF-THE-ART METHODS

In order to evaluate the effectiveness and robustness of MA-Net for liver and tumors segmentation, we also compare the proposed method with other existing state-of-the-art methods. The U-Net [4] is a famous network of encoder-decoder architecture in the medical image segmentation field. Except for the comparison with U-Net, we run experiments with methods of SegNet [39], Res-Unet [40], U-Net++ [7] and U2Net [41], respectively. The U2Net has two versions: U2NetP (4.7MB) and U2Net (176.3MB). The comparison results are shown in Table 2, where we calculated the mean value of Dice, VOE and RVD of all testing CT images. Moreover, Dice, VOE and RVD are based on the mean±standard deviation. We can see that the proposed method achieved the better segmentation performance than other methods for liver and tumors segmentation. Our method achieves

TABLE 2. The quantitative comparison of different variant methods based U-Net. The value of Dice, VOE and RVD are based on the mean standard deviation.

Method	Tumor			Liver		
	Dice	VOE	RVD	Dice	VOE	RVD
U-Net[4]	0.612±0.12	0.51±0.09	-0.23±0.12	0.912±0.04	0.13±0.09	0.17±0.07
SegNet[39]	0.514±0.09	0.62±0.12	-0.48±0.18	0.904±0.02	0.15±0.11	-0.24±0.09
Res-Unet[40]	0.686±0.13	0.48±0.06	0.20±0.09	0.911±0.02	0.11±0.05	-0.15±0.05
U-Net++[7]	0.721±0.11	0.28±0.08	-0.21±0.11	0.938±0.03	0.12±0.08	0.05±0.03
U2NetP[41]	0.713±0.14	0.31±0.07	0.24±0.06	0.936±0.02	0.11±0.06	-0.06±0.04
U2Net[41]	0.768±0.06	0.18±0.08	-0.15±0.04	0.963±0.03	0.04±0.04	0.02±0.02
MA-Net	**0.749±0.08**	**0.21±0.06**	**-0.18±0.07**	**0.960±0.03**	**0.08±0.02**	**-0.03±0.03**

TABLE 3. The segmentation result of other state-of-the-art methods of liver and tumor segmentation on the testing dataset (%).

Method	Dimension	Tumor			Liver		
		Dice	VOE	RVD	Dice	VOE	RVD
Densely FCN[42]	2D	0.625	0.41	19.71	0.923	0.15	-0.08
RC-Unet[43]	2D	0.587	-	-	-	-	-
GIU-Net[44]	2D	-	-	-	0.951	0.11	-0.02
USE-Net[15]	2D	0.741	0.24	-0.19	0.956	0.09	-0.01
H-DenseUNet[6]	3D	0.824	0.36	4.27	0.965	0.07	1.8
U-Net[45]	3D	0.723	-	-	0.946	-	-
RA-Unet[35]	3D	0.795	0.39	-0.15	0.963	0.05	0.02
MA-Net	**2D**	**0.749**	**0.21**	**-0.18**	**0.960**	**0.08**	**-0.03**

the Dice value 0.749±0.08 for tumors segmentation and 0.960±0.03 for liver segmentation. The U2Net has 176.3MB parameters, while the MA-Net only has 19.42MB parameters. The U2Net has about 9 times as many parameters as MA-Net. The U2Net model is more complex than our model. It takes more than three times as long to train model as MA-Net in the same experimental platform. Though having less parameters, MA-Net obtains competitive results for liver and tumor segmentation.

In addition, we list 5 CT images which contains liver and tumor to see the segmentation result visually. These images are selected random in the LiTS dataset. The segmentation results of each method are shown in Fig6. The liver tumor segmentation is seen as the most difficult segmentation task compare to the liver segmentation because the shape of the liver tumor is variable and the size is uncertain. From Fig6, we can see that the proposed method obtains better segmentation performance than other methods for liver and tumors segmentation. There are several deep learning models proposed for liver and tumor segmentation by using the dataset of the LiTS challenges. We compare MA-Net with other state-of-the-art methods. We reach a 0.749 Dice for tumor segmentation and 0.960 Dice for liver segmentation respectively, which is a desirable performance for liver and tumor segmentation. The table 3 lists the detail results and shows all the performances. The MA-Net performs better than 2D networks. The MA-Net is slightly worse than the 3D network for liver and tumor segmentation, such as 3D RA-Unet and 3D H-DenseUNet. While these methods are 3D model, they have more parameters than MA-Net. Meanwhile, their network are significantly complicated than our model. They have longer training time (21 hours on the two NVIDIA Titan Xp platforms such as 3D H-DenseUNet) than MA-Net (8 hours on the one NVIDIA 2070S platform). Moreover, the MA-Net do not use post-processing.

V. DISCUSSION

Automatic liver and tumors segmentation is helpful for radiologists in clinical diagnosis, which provides the precise contour of the liver and tumors for radiologists and assists radiologists in clinical process. We design a novel network based on improved U-Net for liver and tumors segmentation. We introduce self-attention mechanism into the proposed method which contains two blocks of self-attention mechanism, PAB and MFAB. PAB considers the spatial dependencies between pixels, and MFAB considers the channel dependencies between any feature maps.

To verify the effectiveness and robustness of the proposed method, the ablation analysis is used firstly in this experiment. By the ablation analysis, the PAB and MFAB are effective for liver and tumors segmentation. In order to eliminate the influence of increasing parameter for the segmentation

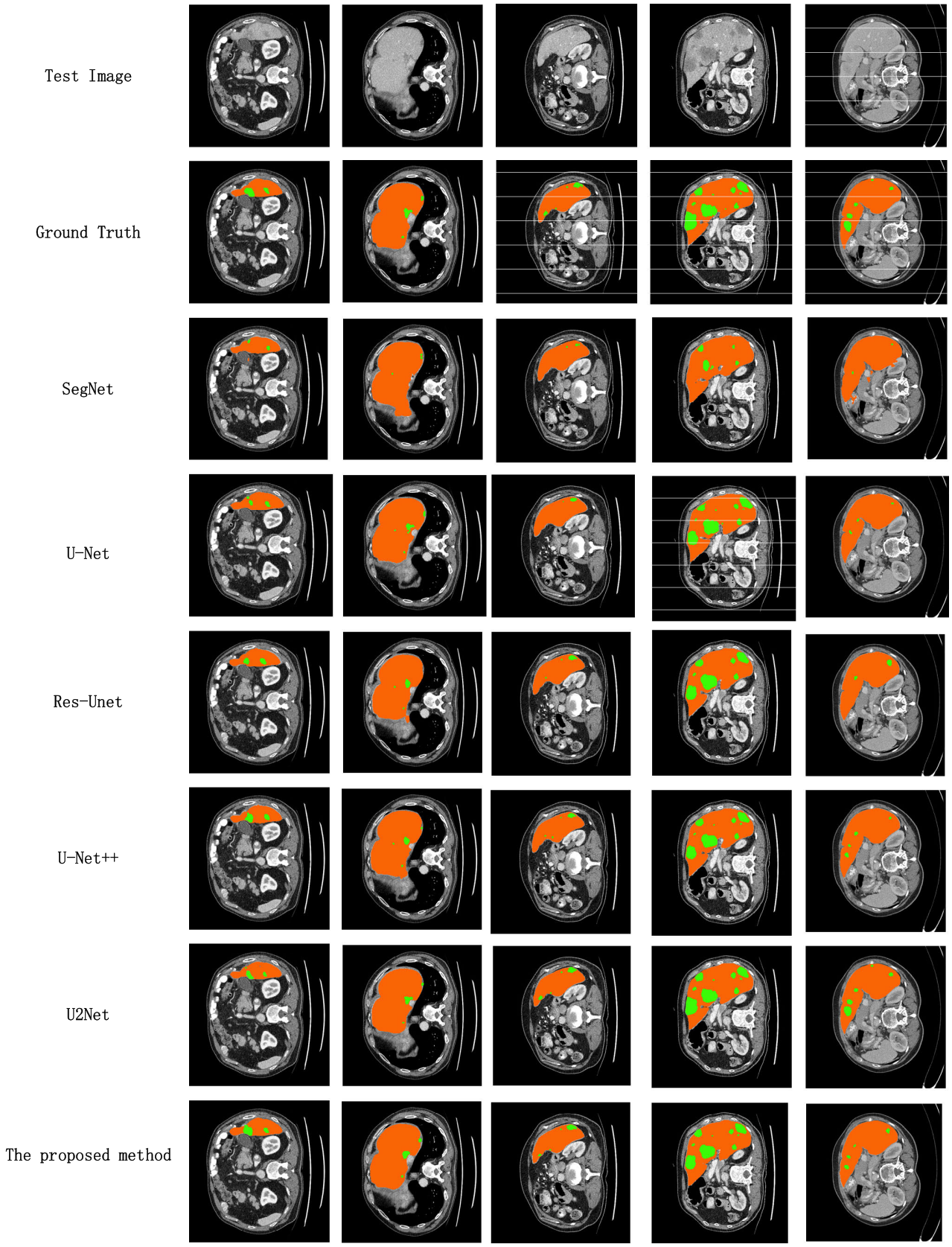


FIGURE 6. Examples of liver and tumors segmentation results of the different methods on the testing dataset. The orange region denotes the livers and the green region denotes the tumors.

accuracy, we compare the proposed method with the U-Net+wide which has the same parameter number. That also shows the superiority of our method.

To further prove the superiority of the proposed method, we compare MA-Net with other state-of-the-art methods for liver and tumor segmentation on testing dataset. Table 2 shows the results compared to other methods. We can see that the segmentation performance of the proposed method outperforms other methods. The proposed method achieves 0.960±0.03 (Dice) and 0.749±0.08 (Dice) on the liver and tumors segmentation via measuring the Dice value respectively. Moreover, we list some Examples of liver and tumors segmentation results on the testing dataset as shown in Fig6. The tumors segmentation is a challenging task compared to liver segmentation. Fig6 shows that the proposed method still performs well for tumors segmentation. The application of attention mechanism contributes to further improve the segmentation performance in the CNNs. The proposed method can provide accurate guidance for doctor in clinical diagnosis. Compare to some other state-of-the-art methods of 2D and 3D U-Net models, our model obtains better segmentation results. While MA-Net is slightly worse than some 3D models, the MA-Net has fewer parameters and simpler model than it. For liver segmentation, the MA-Net performs better than most 2D networks and is comparable to some 3D models. For tumor segmentation, the MA-Net performs better than most 2D networks. The MA-Net can also provide good guidance to doctors.

VI. CONCLUSION

We design a novel network architecture based on improve U-Net for liver and tumors segmentation. We introduce self-attention mechanism into our method to segmentation image. Specially, we use self-attention mechanism to capture the spatial and channel dependencies of feature maps and consider the Multi-scale semantic information based on the channel dependencies between any feature maps. In addition, we use a new loss function which combined the cross entropy and Dice.

Massive experiments demonstrated the superiority of our method on the 2017 LiTS dataset. The proposed method is helpful to assist the doctor in clinical process. However, the MA-Net also has some shortcomings and it just can segmented the liver and tumor in this paper. We will study the effect of MA-Net on other medical images to assess the segmentation performance and robustness of MA-Net in future studies. Moreover, we consider adding deep supervision into MA-Net to improve model. However, we mainly focus on introducing attention mechanism into medical image segmentation filed and the MA-Net not apply the 3D information of CT images. The 3D information of CT images is also important for liver and tumor segmentation. We will consider adding 3D information of CT images to the MA-Net model in the future studies. The model combines 3D information of CT images and attention mechanism to further improves model. The MA-Net has very good potential for further development.

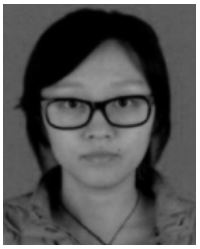
REFERENCES

[1] J. Ferlay, H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin, "Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008," *Int. J. Cancer*, vol. 127, no. 12, pp. 2893–2917, Dec. 2010.
[2] R. Lu, P. Marziliano, and C. Hua Thng, "Liver tumor volume estimation by semi-automatic segmentation method," in *Proc. IEEE Eng. Med. Biol. 27th Annu. Conf.*, Jan. 2006, pp. 3296–3299.
[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[6] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE Trans. Med. Imag.*, vol. 37, no. 12, pp. 2663–2674, Dec. 2018.
[7] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[9] J. Ni, J. Wu, J. Tong, Z. Chen, and J. Zhao, "GC-net: Global context network for medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105121.
[10] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Trophectoderm segmentation in human embryo images via inceptioned U-Net," *Med. Image Anal.*, vol. 62, May 2020, Art. no. 101612.
[11] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[12] X. Fang and P. Yan, "Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction," 2020, *arXiv:2001.00208*. [Online]. Available: http://arxiv.org/abs/2001.00208
[13] T. Song, F. Meng, A. Rodriguez-Paton, P. Li, P. Zheng, and X. Wang, "U-next: A novel convolution neural network with an aggregation U-Net architecture for gallstone segmentation in CT images," *IEEE Access*, vol. 7, pp. 166823–166832, 2019.
[14] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, and J. Liu, "CE-net: Context encoder network for 2D medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2281–2292, Oct. 2019.
[15] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M. S. Nobile, C. Ferretti, D. Besozzi, M. C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama, and P. Cazzaniga, "USE-net: Incorporating squeeze-and-excitation blocks into U-net for prostate zonal segmentation of multi-institutional MRI datasets," *Neurocomputing*, vol. 365, pp. 31–43, Nov. 2019.
[16] J. Zhang, Y. Jin, J. Xu, X. Xu, and Y. Zhang, "MDU-net: Multi-scale densely connected U-net for biomedical image segmentation," 2018, *arXiv:1812.00352*. [Online]. Available: http://arxiv.org/abs/1812.00352
[17] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, and X. Cao, "Joint optic disc and cup segmentation based on multi-label deep network and polar transformation," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1597–1605, Jul. 2018.
[18] S. Zhou, D. Nie, E. Adeli, J. Yin, J. Lian, and D. Shen, "High-resolution Encoder–Decoder networks for low-contrast medical image segmentation," *IEEE Trans. Image Process.*, vol. 29, pp. 461–475, 2020.
[19] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, Jan. 2020.
[20] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3D semantic scene completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7801–7810.
[21] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, and C. Zhang, "Disan: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–10.

[22] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3194–3203.
[23] Z. Lin, M. Feng, C. Nogueira dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*. [Online]. Available: http://arxiv.org/abs/1703.03130
[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[25] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[26] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[27] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 267–283.
[28] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9167–9176.
[29] H. Hu, Z. Zhang, Z. Xie, and S. Lin, "Local relation networks for image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3464–3473.
[30] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318
[31] Y. Zhou, W. Huang, P. Dong, Y. Xia, and S. Wang, "D-UNet: A dimension-fusion u shape network for chronic stroke lesion segmentation," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Sep. 6, 2019, doi: 10.1109/TCBB.2019.2939522.
[32] Y. Wang, D. Ni, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, J. Qin, P.-A. Heng, and T. Wang, "Deep attentive features for prostate segmentation in 3D transrectal ultrasound," *IEEE Trans. Med. Imag.*, vol. 38, no. 12, pp. 2768–2778, Dec. 2019.
[33] S. Li, G. K. F. Tso, and K. He, "Bottleneck feature supervised U-Net for pixel-wise liver and tumor segmentation," *Expert Syst. Appl.*, vol. 145, May 2020, Art. no. 113131.
[34] Z. Liu, Y.-Q. Song, V. S. Sheng, L. Wang, R. Jiang, X. Zhang, and D. Yuan, "Liver CT sequence segmentation based with improved U-Net and graph cut," *Expert Syst. Appl.*, vol. 126, pp. 54–63, Jul. 2019.
[35] Q. Jin, Z. Meng, C. Sun, L. Wei, and R. Su, "RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans," 2018, *arXiv:1811.01328*. [Online]. Available: https://arxiv.org/abs/1811.01328
[36] Y. Wu and K. He, "Group normalization," 2018, *arXiv:1803.08494*. [Online]. Available: http://arxiv.org/abs/1803.08494
[37] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.
[38] P. Bilic *et al.*, "The liver tumor segmentation benchmark (LiTS)," 2019, *arXiv:1901.04056*. [Online]. Available: http://arxiv.org/abs/1901.04056
[39] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[40] X. Han, "Automatic liver lesion segmentation using a deep convolutional neural network method," 2017, *arXiv:1704.07239*. [Online]. Available: http://arxiv.org/abs/1704.07239
[41] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested U-structure for salient object detection," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.
[42] K. Chaitanya Kaluva, M. Khened, A. Kori, and G. Krishnamurthi, "2D-densely connected convolution neural networks for automatic liver and tumor segmentation," 2018, *arXiv:1802.02182*. [Online]. Available: http://arxiv.org/abs/1802.02182
[43] R. K. Pandey, A. Vasan, and A. G. Ramakrishnan, "Segmentation of liver lesions with reduced complexity deep models," 2018, *arXiv:1805.09233*. [Online]. Available: http://arxiv.org/abs/1805.09233
[44] Z. Liu, Y.-Q. Song, V. S. Sheng, L. Wang, R. Jiang, X. Zhang, and D. Yuan, "Liver CT sequence segmentation based with improved U-Net and graph cut," *Expert Syst. Appl.*, vol. 126, pp. 54–63, Jul. 2019.
[45] S. Rafiei, E. Nasr-Esfahani, K. Najarian, N. Karimi, S. Samavi, and S. M. R. Soroushmehr, "Liver segmentation in CT images using three dimensional to two dimensional fully convolutional network," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2067–2071.

TONGLE FAN was born in Baoding, China, in 1994. He is currently pursuing the M.A.Eng. degree with Hebei University. His research interests include artificial intelligence and medical image processing.

GUANGLEI WANG was born in Tianjin, China, in 1983. He received the B.E. degree in precision instrumentation and opto-electronics from Tianjin University, China, the M.A.Eng. degree in biomedical engineering from the Chalmers University of Technology, Sweden, and the Ph.D. degree in mechanical engineering from the University of Padova, Italy. His research interests include medical image processing and biomechanics.



YAN LI was born in Tonglin, China, in 1985. She received the B.E. degree from the University of Hertfordshire, U.K., the MA.Eng. degree in electrical and electronics from the Chalmers University of Technology, Sweden, and the Ph.D. degree in biomedical engineering from the Cluj-Napoca University of Technology, Italy, in 2013. Her research interests include image analysis and pattern recognition.

HONGRUI WANG was born in Keshan, China, in 1956. He received the M.A.Eng. and Ph.D. degrees in industrial automation from Yanshan University, China. His research interests include parallel control theory and computer.