Mathematical Biosciences and Engineering

MBE, 20(1): 1297–1316.
DOI: 10.3934/mbe.2023059
Received: 19 September 2022
Revised: 22 October 2022
Accepted: 24 October 2022
Published: 27 October 2022

http://www.aimspress.com/journal/MBE

Research article

Multi-scale attention and deep supervision-based 3D UNet for automatic liver segmentation from CT

Jinke Wang[1,2,*], Xiangyang Zhang[2], Liang Guo[2], Changfa Shi[3] and Shinichi Tamura[4]

[1] Department of Software Engineering, Harbin University of Science and Technology, Rongcheng 264300, China
[2] School of Automation, Harbin University of Science and Technology, Harbin 150080, China
[3] Mobile E-business Collaborative Innovation Center of Hunan Province, Hunan University of Technology and Business, Changsha 410205, China
[4] SANKEN, Osaka University, Suita 565-0871, Japan

* **Correspondence:** Email: jkwang@hitwh.edu.cn; Tel: +8613863132787; Fax: +8606317595512.

Abstract: *Background*: Automatic liver segmentation is a prerequisite for hepatoma treatment; however, the low accuracy and stability hinder its clinical application. To alleviate this limitation, we deeply mine the context information of different scales and combine it with deep supervision to improve the accuracy of liver segmentation in this paper. *Methods*: We proposed a new network called MAD-UNet for automatic liver segmentation from CT. It is grounded in the 3D UNet and leverages multi-scale attention and deep supervision mechanisms. In the encoder, the downsampling pooling in 3D UNet is replaced by convolution to alleviate the loss of feature information. Meanwhile, the residual module is introduced to avoid gradient vanishment. Besides, we use the long-short skip connections (LSSC) to replace the ordinary skip connections to preserve more edge detail. In the decoder, the features of different scales are aggregated, and the attention module is employed to capture the spatial context information. Moreover, we utilized the deep supervision mechanism to improve the learning ability on deep and shallow information. *Results*: We evaluated the proposed method on three public datasets, including, LiTS17, SLiver07, and 3DIRCADb, and obtained Dice scores of 0.9727, 0.9752, and 0.9691 for liver segmentation, respectively, which outperform the other state-of-the-art (SOTA) methods. *Conclusions*: Both qualitative and quantitative experimental results demonstrate that the proposed method can make full use of the feature information of different stages while enhancing spatial data's learning ability, thereby achieving high liver segmentation accuracy. Thus, it proved to be a promising tool for automatic liver segmentation in clinical assistance.

Keywords: liver segmentation; attention; deep supervision; CT; deep learning

1. Introduction

Liver segmentation from abdominal CT plays an essential role in various clinical applications. However, radiologists still predominantly perform this task in a slice-by-slice fashion, which is labor-intensive and prone to errors due to observer dependence. Therefore, automatic and accurate liver segmentation technology is highly desirable in the clinical environment.

Currently, automatic liver segmentation methods can be divided into classical machine learning-based and deep learning-based approaches. The former mainly includes thresholding [1], region growing [2], superpixel [3], level sets [4], sparse [5], atlas [6], etc. However, although machine learning-based methods significantly improved the segmentation accuracy, they still require artificial feature engineering intervention, resulting in unsatisfactory robustness.

Thanks to its remarkable feature learning ability, the deep learning-based method has attracted many scholars to the medical image process. Long et al. [7] first proposed the fully convolutional networks (FCN), which replaced the fully connected layer of VGG16 [8] with a convolutional layer. They restored the image to the original resolution through deconvolution, realizing the pixel-level prediction. Then, Ronneberger et al. [9] proposed the U-Net with a fully symmetric encoder and decoder based on FCN, which can obtain more refined results through gradual upsampling. Due to its excellent performance in medical image segmentation, scholars have successively developed various improved methods, including three categories: 1) 2D-based, 2) 3D-based, and 3) 2.5D-based methods.

The 2D-based methods require the least memory. Liu et al. [10] introduced the residual module [11] into U-Net and designed a cascaded liver segmentation model to alleviate the gradient vanishment. Xi et al. [12] proposed U-ResNets for liver and tumor segmentation. To address the imbalance issue of image category, they evaluated the model with five different loss functions. Oktay et al. [13] proposed Attention U-Net, which adds the attention gate to the skip connection of UNet. The attention gate can automatically distinguish the shape and size of the target so that the network pays more attention to the area of interest while suppressing the irrelevant area. Hong et al. [14] proposed the quartet attention UNet (QAUNet). They use quartet attention to capture the intrinsic and cross-dimensional features between channels and spatial locations. They verified the effectiveness of the network in segmenting liver and tumor through extensive experiments. Finally, Cao et al. [15] suggested a dual-attention model for liver tumor segmentation and introduced an attention gate into DenseUNet to reduce the response of irrelevant regions. In addition, the attention in the bidirectional Long Short Term Memory (LSTM) appropriately adjusts the weights of the two types of features according to their contributions to the improvement of encoding and upsampling.

For 3D-based approaches, Ji et al. [16] developed a 3D convolutional neural network (CNN) that extracts features from both spatial and temporal dimensions via 3D convolutions. Based on U-Net and 3D CNN, Cicek et al. [17] proposed 3D UNet, which replaced all 2D operations with 3D processes. Milletari et al. [18] proposed VNet. It deepened the network's depth, replaced the downsampling pooling with convolution, and achieved superior performance compared to 3D UNet. In addition, Liu et al. [19] proposed an improved 3D UNet combined with graph cutting for liver segmentation. Lei et al. [20] designed a lightweight VNet. During the training phase, they employed 3D deep supervision to improve the loss function, which showed great discriminative

ability in dealing with liver and non-liver regions. Zhou et al. [21] proposed a novel memory-augmented network, the Volumetric Memory Network (VMN), for interactive segmentation of volumetric medical data. It solves the task by sequential label propagation while considering the rich 3D structures, thus avoiding costly 3D operations. Extensive experiments showed superior performance compared with a reasonable number of user interactions. Finally, Jin et al. [22] proposed a 3D hybrid residual attention-aware segmentation approach, which combines low- and high-level feature information and achieves Dice of 0.961/0.977 for liver segmentation on LiTS17/3DIRCADb datasets, respectively.

The 2.5D-based methods can significantly reduce the memory requirement by utilizing part of the inter-slice information of 3D data. Han et al. [23] developed a deep CNN that takes the stack of adjacent slices as input and generates a segmentation map corresponding to the central slice, realizing the 2.5D mode of the network. Li et al. [24] proposed H-DenseUNet based on 2D and 3D intra- and inter-slice information for liver and liver tumor segmentation. The network first extracts the image information through the 2D network. It then associates the pixel probability generated by the 2D network with the original 3D volume. Lv et al. [25] proposed a 2.5D light liver segmentation network. They leverage the techniques from the residual and Inception theories, reducing the number of parameters by 70% compared with UNet.

Nevertheless, each of the methods mentioned above cannot be used straightforwardly to generate a satisfactory result in certain challenging cases, which can be outlined as follows: (i) There are other issues around the liver or organs with similar intensity; (ii) There are multiple discrete small liver regions; (iii) The edge of the liver contains tumors.

To effectively alleviate the above issues, we developed an end-to-end 3D network framework, MAD-UNet, to aggregate multi-scale attention and combined it with deep supervision. The main contributions are summarized as follows:

- Use LSSC to avoid redundant processing of low-resolution information and improve the feature fusion of low- and high-resolution information.
- Employ attention mechanism to aggregate multi-scale features, making full use of the contextual spatial information at different scales.
- Combine the binary cross-entropy loss with Dice loss, and apply deep supervision to the features of different levels to improve the accuracy.
- Validate the proposed method on three publicly available datasets.

The rest of this paper is organized as follows. Section 2 introduces the related work; Section 3 describes the proposed network framework; Section 4 gives the experimental results and analysis in detail, and the last section provides the conclusions of this paper.

2. Materials and methods

2.1. Attention mechanism

Since ordinary convolution operations often failed to highlight the target features and suppress the hidden layer's noise, the attention *mechanism* was proposed and found to be an effective model for alleviating such problems. For example, Squeeze-and-Excitation (SE) Block [26] could optimize the quality of representations by modeling the inter-dependencies between convolutional feature channels and thus significantly improve the performance of existing networks at a slightly increased

computational cost. Woo et al. [27] developed the Convolutional Block Attention Module (CBAM). Given an intermediate feature map, the module first infers attention maps sequentially along two separate dimensions, channel and spatial. Then, the attention maps are multiplied to the input feature map for adaptive feature refinement. The CBAM can also be seamlessly integrated into any CNN architecture.

In addition, Li et al. [28] applied 3D channel attention and 3D spatial attention modules in the decoder to extract features from different scales and achieve competitive performance in spine segmentation. Zhou et al. [29] proposed a novel Motion-Attentive Transition Network (MATNet) for zero-shot video object segmentation. They designed an asymmetric attention block called Motion-Attentive Transition (MAT) in a two-stream encoder, which can convert appearance features to motion-attentive representations at each convolution stage. This design has the benefit of allowing the encoder to be deeply interleaved and to allow a tight hierarchical interaction between object motion and appearance. Wang et al. [30] adopted a multi-resolution attention module to combine local deep attention features (DAF) with global background for prostate segmentation on ultrasound images. They combined the local and global features in a simple attention module and then produced an attention map through the sigmoid function to model long-range dependencies.

2.2. Deep supervision mechanism

Lee et al. [31] first proposed the deeply supervised network. They improve CNN's convergence speed and recognition ability in image classification by supervising the training of hidden layers. For medical image segmentation, effectively segmenting organs in volumetric images requires deep networks to extract features. However, training a deep network may cause gradient vanishment or explosion problems, resulting in ineffective backpropagation of loss.

To address this issue, Dou et al. [32] proposed to utilize direct supervision to train hidden layers in 3D FCN. They use deconvolution to upscale the low- and mid-level features and then exert a softmax function on these upscaled features to obtain ultra-dense predictions. Finally, they calculated the classification errors of the prediction results of these branches and ground truth and verified the effectiveness of deep supervision on the SLiver07 dataset. Wang et al. [33] introduced deep supervision into 3D FCNs. It effectively alleviated the gradient exploding and vanishing problem, which is commonly encountered in deep model training, thereby forcing the update process of hidden layer filters to be conducive to high-resolution features. Yang et al. [34] developed a dual-path deep supervision mechanism. One is to generate multiple predictions from multiple semantic layers and average them to produce accurate segmentation. The other is to adjust the weight of the layer by monitoring the local depth of the learned features. Their deeply supervised approach achieves good performance in lung tumor segmentation.

3. Methods

3.1. Proposed network *framework*

The proposed MAD-UNet[1] network framework is shown in Figure 1. The overall framework consists of 3D UNet, multiple attention modules, and deep supervision operations. In the encoder,

[1] Our source code is available at https://github.com/ZhangXY-123/Model/blob/master/MAD-UNet.py

ordinary convolutions are replaced by residual blocks to prevent gradients vanishment. In the downsampling, to retain more feature information, convolution with a kernel size of 3 × 3 × 3 and a stride of 2 is used to replace the pooling operation. In the skip connection stage, LSSC (Figure 2) is used instead of ordinary skip connections to avoid the repetition of low-resolution feature information. In the decoder, the number of features channels before each upsampling is first halved to reduce the number of parameters. At the same time, the features of different resolutions of the decoder are upsampled to the same resolution size to form multiple SLFs (single-layer feature). Then MLFs (Multi-Layer Feature) are formed through splicing and convolution. Finally, the multi-scale fusion of MLF and SLF maps of different resolutions is used to extract regions of interest through the attention module (Figure 3), and multiple AFM (Attentive Feature Maps) are obtained. Then the liver segmentation is generated by concatenation, convolution, and the Sigmoid activation on these four AFMs. In this process, the small-scale feature maps have low resolution but a high level of semantic information. On the contrary, the large-scale feature maps have high resolution but rich details. MLFs are used to effectively deal with liver regions of different sizes and complex shapes. And four SLFs, four AFMs, and the final output image are deeply supervised.

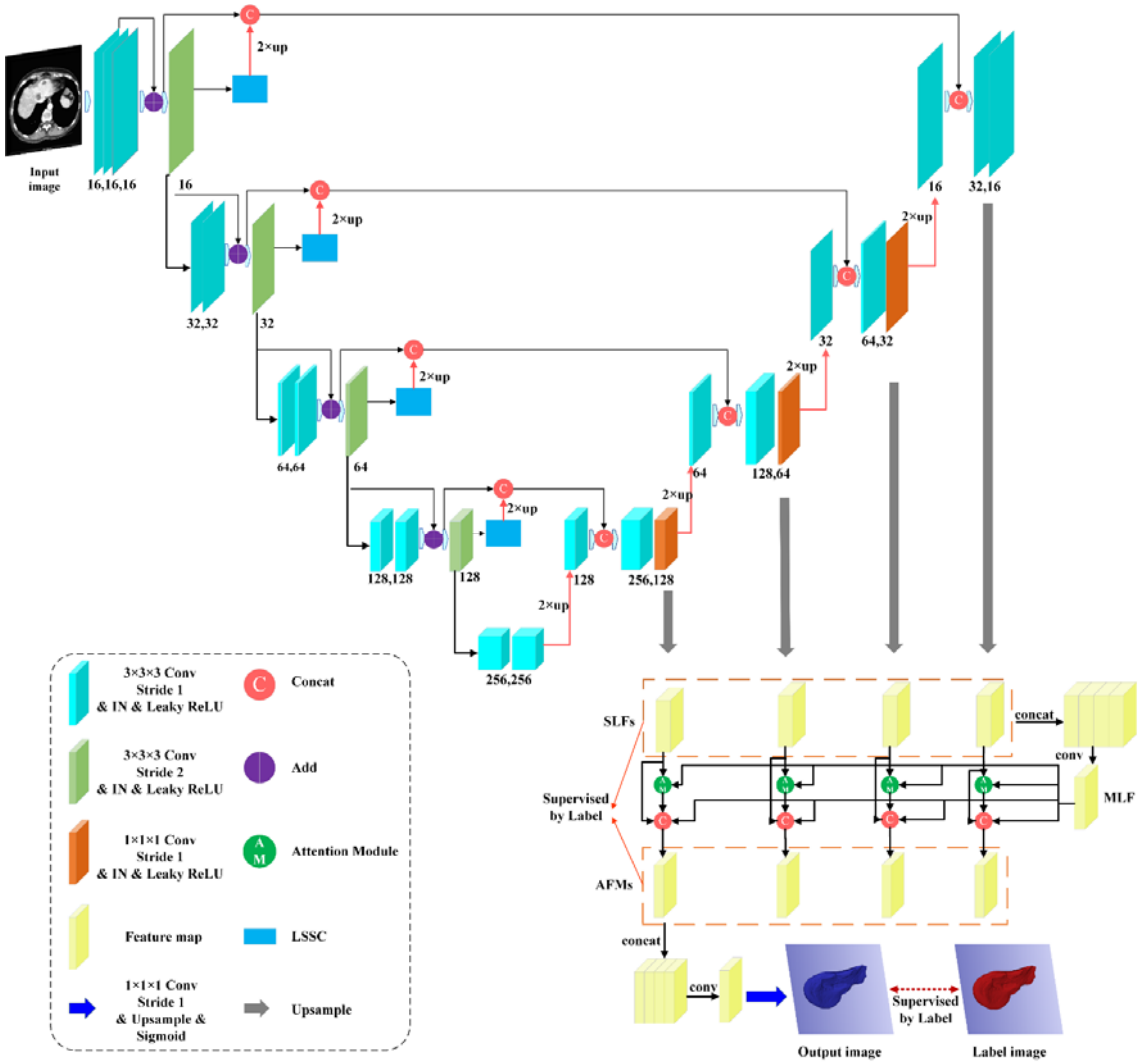

Figure 1. The proposed MAD-UNet framework.

3.2. Long-short skip connection

The convolutional and pooling operations in U-Net can obtain deeper semantic features while reducing the image resolution. However, pooling often hinders the downward transmission of shallow features such as edges, resulting in most low-resolution semantic features being transmitted without enough edge information or small target features. Therefore, in this paper, we employed LSSC to improve this part of the problem.

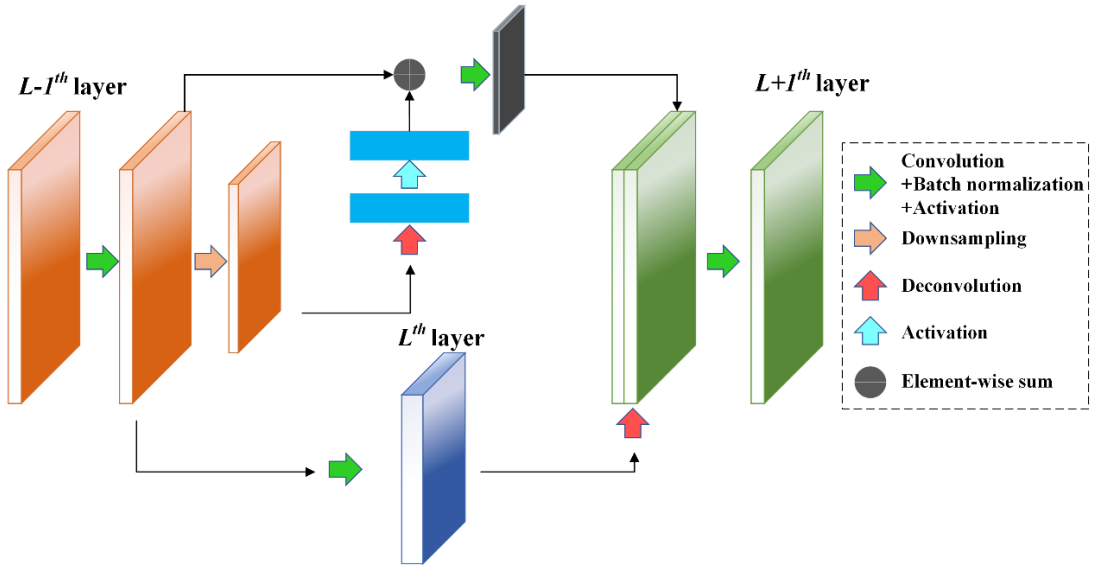

Figure 2. The structure of the LSSC.

In Figure 2, we detailed the LSSC modules of the transition stage between the encoder and the decoder. "L^{th}" represents the transition layer between the encoder and decoder, and "$L\text{-}1^{th}$" is the last transition layer, while "$L+1^{th}$" represents the next level of the transition layer. We use a residual module to avoid the repetition of the low-resolution feature. The residual module consists of deconvolution and an activation function. It is added after downsampling. The deconvolution of the residual path is first restored to the feature size before downsampling. Then, the obtained target features are directly passed to the decoder through skip connections. To help the network obtain edge features that are ignored in ordinary skip connections, an extra set of convolution blocks is added to each skip connection, consisting of a convolution, a batch normalization, and an activation function. Compared with ordinary skip connections, LSSC can effectively retain the edge features of the target and meanwhile avoid repeated input features.

3.3. Attention module

Attention mechanisms have been applied in various image-processing tasks. For example, SENet [26] improves the representative ability of the network by establishing interdependencies between convolutional feature channels. CBAM [27] fuses channel attention and spatial attention, enabling seamless integration into any CNN network.

In the task of liver segmentation, the shallow feature map contains detailed information about

the liver, and many non-liver regions as well. On the other hand, the deep feature map can obtain semantic details on the location information of the liver but may lose the points of the liver edge. To refine the information of each layer, we employed the deep attention module to generate refined features. The proposed attention module explores the effect of the hierarchical attention mechanism in liver segmentation, selectively using complementary features at all scales to refine features at different levels, thereby boosting segmentation accuracy. Its structure is shown in Figure 3.

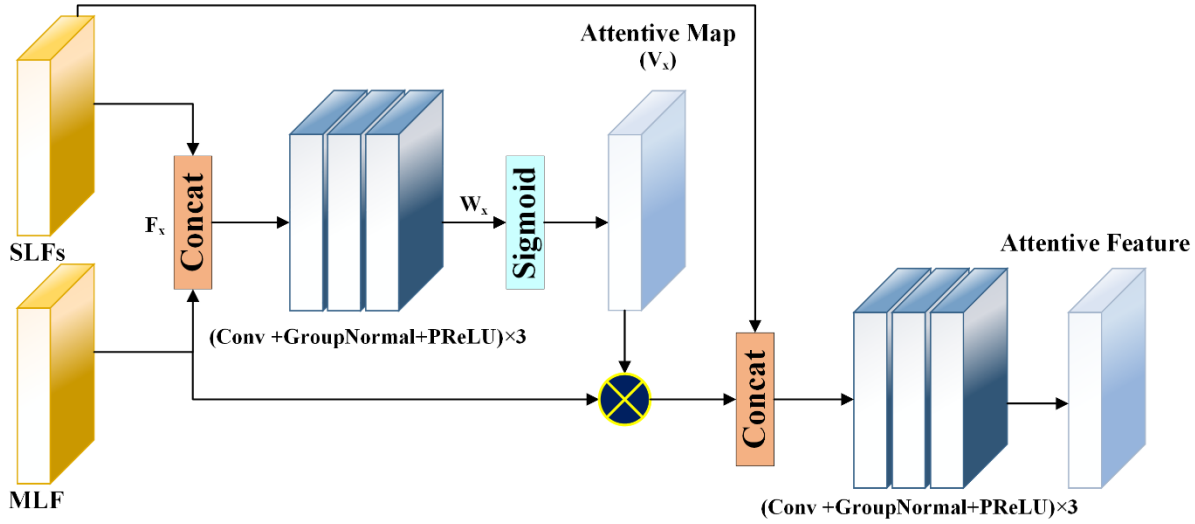

Figure 3. The structure of the attention module.

First, the SLF and the MLF are concatenated as F_x, and then the attention weight W_x is generated by 3 f_a operations,

$$W_x = f_a(F_x; \theta)$$

where θ represents the learning parameters of f_a, which contains three convolution layers consisting of two 3 × 3 × 3 convolutions and one 1 × 1 × 1 convolution. Each convolutional layer consists of one convolutional, one group normalization, and one PReLU. These convolution operations can select useful multi-level information according to the features of a single layer. The attention module computes the attentive map A_x by normalizing W_x with a Sigmoid function. Next, multiply the attention map by MLF to weigh the features in the MLF of each SLF. Finally, the weighted MLFs are merged with the corresponding features of each SLF by applying Conv + GroupNormal + PReLU again, which can automatically refine the SLF layer by layer and generate a given layer's final attention features. In this way, we can simultaneously utilize the advantages of SLFs and MLFs. Specifically, it suppresses the detailed information not in the semantically salient region, captures more details in the semantically salient region, and enhances the boundary details.

3.4. Loss function

The cross-entropy loss function is commonly used for segmentation tasks since it can well retain boundary information; however, it is prone to produce significant errors when dealing with cost imbalance problems. In contrast, Dice loss has good performance for scenes with severely imbalanced

positive and negative samples. Still, training loss would show instability in processing small targets. Therefore, we utilize a combination of Dice loss and binary cross-entropy loss to take into account the similarity of local details and global shapes, which are defined in Eqs (1) and (2), respectively.

$$L_{BCE} = \sum_{i=1}^{N} g_i \log p_i + \sum_{i=1}^{N} (1 - g_i) \log(1 - p_i) \tag{1}$$

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i^2 + \sum_{i=1}^{N} g_i^2} \tag{2}$$

where N is the voxel number of the input volume; $p_i \in [0.0,1.0]$ represents the voxel value of the predicted probabilities; $g_i \in \{0,1\}$ is the voxel value of the binary ground truth volume.

In the process of training, the supervision signal of each stage adopts the combination of L_{BCE} and L_{Dice}, which are defined in Eq (3). We utilize a total of nine deep supervision signals and the L_{total} is defined as the sum of all signals, defined in Eq (4), in which, w^i and L_{signal}^irepresent the weight and loss of the *i*-th layer, respectively, while w^j and L_{signal}^j represent the weight and loss of the *j*-th layer after the features refinement by the attention module, *n* represents the number of layers of the network, and w^f and L_{signal}^f denote the weight and loss of the output layer, respectively. In this paper, we empirically set the weights $w^{i=1,2,3,4}$, $w^{j=1,2,3,4}$and w^f as (0.2, 0.4, 0.6, 0.8), (0.3, 0.5, 0.7, 0.9) and 1, respectively.

$$L_{signal} = L_{BCE} + L_{Dice} \tag{3}$$

$$L_{total} = \sum_{i=1}^{n} w^i L_{signal}^i + \sum_{j=1}^{n} w^j L_{signal}^j + w^f L_{signal}^f \tag{4}$$

3.5. Evaluation metrics

In this experiment, we choose five metrics to evaluate the performance of the proposed method, including Dice, volumetric overlap error (VOE), relative volume difference (RVD), average symmetric surface distance (ASD), and root mean square symmetric surface distance (RMSD) [35].

In addition, to validate whether the difference in segmentation accuracy between our proposed method and the comparison methods was statistically significant, we performed paired *t*-tests on two key metrics (Dice and ASD) with a significance level of $p < 0.05$. The null hypothesis is that the mean values of the same evaluation metric are the same for the compared methods.

4. Experiments and results

4.1. Datasets and implementation

We tested the proposed method on three public datasets: LiTS17[2], SLiver07[3], and 3DIRCADb[4]. For the LiTS17 dataset, we randomly select 116 sets of data for training (3:1) and 15 sets for testing. For both SLiver07 and 3DIRCADb datasets, we randomly choose 12 sets of data for training and eight groups for testing. The details of the three datasets are listed in Table 1.

[2] The dataset is publicly available at https://competitions.codalab.org/competitions/17094#results
[3] The dataset is publicly available at https://sliver07.grand-challenge.org
[4] The dataset is publicly available at https://www.ircad.fr/research/3d-ircadb-01/

Table 1. Detailed parameters of the three public liver CT datasets.

Datasets	Total train set	In-plane resolution	Inter-slice spacing	Slice num	Size
LiTS17	131	0.55–1.0 mm	0.45–6.0 mm	75–987	512 × 512
SLiver07	20	0.5–0.8 mm	1.0–3.0 mm	64–394	512 × 512
3DIRCADb	20	0.56–0.81 mm	1.0–4.0 mm	74–260	512 × 512

To reduce the training time and improve the computational efficiency, we set the volume of the input image to 16 × 256 × 256. Besides, to exclude irrelevant organs, we adjusted the greyscale to [-200, 200] HU by windowing process, set the z-axis spacing of all data was to 1 mm and removed the slices without liver were. Finally, we expanded 20 slices to the front and back of the area containing the liver.

In addition, we chose Adam as the optimizer in the training process and combined BCE loss with Dice loss as the loss function. We set the initial learning rate to 0.001, which would update according to $lr = initial\_lr \times \gamma$. When the epoch reaches 400/650, the learning rate starts to decay, and the initial value of γ is set to 0.1. A total of 800 epochs were trained with a batch size of 1. We run the experiments on a workstation with Ubuntu 18.04, graphics card RTX2080Ti, RAM 32G, single CPU Intel Xeon Silver 4110, and Pytorch1.8.

4.2. Ablation experiment

To verify the effectiveness of the proposed model, we performed ablation experiments on the LiTS17 dataset. Taking 3D UNet as the baseline, we conducted qualitative comparative experiments with Baseline + LSSC, Baseline + LSSC + Multi-scale Attention (MA), Baseline + LSSC + MA + Deep Supervision (DS).

Table 2. Ablation experiments on the LiTS17 dataset.

Method	Dice (%)	VOE (%)	RVD (%)	ASD (mm)	RMSD (mm)
3D UNet (Baseline)	92.49 ± 5.34*	13.43 ± 9.77	1.05 ± 0.68	2.87 ± 1.43*	8.53 ± 10.78
+LSSC	95.75 ± 1.59*	8.26 ± 2.91	0.73 ± 0.58	1.21 ± 0.88*	5.43 ± 4.76
+LSSC+MA	96.42 ± 1.53*	7.56 ± 2.78	0.54 ± 0.31	1.17 ± 0.59*	4.95 ± 5.23
+LSSC+MA+DS	**97.27 ± 1.22**	**6.83 ± 2.31**	**0.34 ± 0.19**	**1.03 ± 0.37**	**3.74 ± 3.58**

*Note: Bold font represents the best results. * indicates a statistically significant difference between the labeled results and the corresponding results of our method at a significance level of 0.05.

From Table 2, we can see that Baseline + LSSC achieves 0.9575 on Dice, which is 3.26% higher than the Baseline 3D UNet. By introducing a multi-scale attention module, the Dice score of Baseline + LSSC + MA achieved a 0.67% improvement (0.9642). Furthermore, our proposed MAD-UNet superimposed with deep supervision improves the Dice score by 0.85% (0.9727). Besides, the MAD-UNet also resulted in the best score on the other four evaluation metrics and thus proved the effectiveness of the proposed network.

4.3. Test on the LiTS17 dataset

4.3.1. Quantitative comparison on the LiTS17 dataset

To verify the high accuracy of the proposed method, we compared the proposed method with four SOTA methods, including 3D UNet [17], VNet [18], 3D ResUNet [36], and 3D DenseUNet [24].

Figure 4 shows the Dice and Loss values of the five models during training on the LiTS17 dataset. For example, from Figure 4(a), we can see that the Dice of 3D UNet converge the slowest and has the lowest Dice value. In addition, the value of Dice of 3D ResUNet has been the highest in the first 400 epochs. However, after 400 and 650 epochs, as the learning rate decreases, its score is gradually exceeded by the proposed MAD-UNet. Finally, the proposed model outperforms the other four on Dice. In addition, during the training process, the loss of the proposed MAD-UNet (Figure 4(b)) converges in the lowest position.

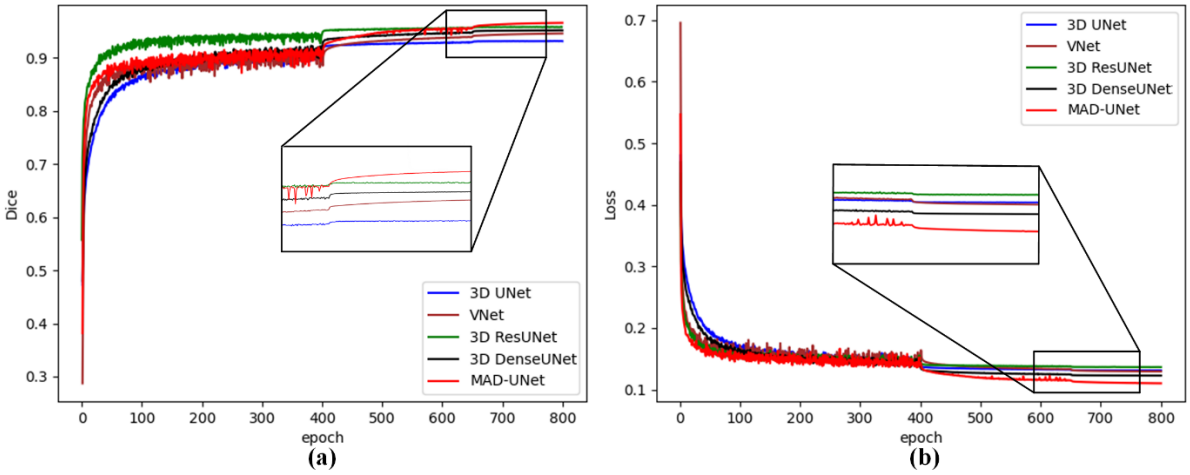

Figure 4. Dice and Loss of different methods during training on the LiTS17 dataset. (a) Dice (b) Loss.

Table 3. Comparative results of different methods on LiTS17 dataset.

Method	Dice (%)	VOE (%)	RVD (%)	ASD (mm)	RMSD (mm)
3D UNet [17]	92.49 ± 5.34*	13.43 ± 9.77	1.05 ± 0.68	2.87 ± 1.43*	8.53 ± 10.78
VNet [18]	93.46 ± 3.03*	12.13 ± 5.23	**-0.18 ± 0.55**	2.45 ± 1.96*	6.45 ± 5.93
3D ResUNet [36]	95.43 ± 2.04*	8.57 ± 3.29	0.29 ± 0.37	1.39 ± 0.92*	4.66 ± 4.78
3D DenseUNet [24]	94.85 ± 2.67*	9.42 ± 4.68	-1.23 ± 0.75	1.46 ± 1.33*	5.32 ± 4.39
Our MAD-UNet	**97.27 ± 1.22**	**6.83 ± 2.31**	0.34 ± 0.19	**1.03 ± 0.37**	**3.74 ± 3.58**

*Note: Bold font represents the best results. * indicates a statistically significant difference between the labeled results and the corresponding results of our method at a significance level of 0.05.

Table 3 shows the comparison with the segmentation results of the SOTA methods. On the LiTS17 dataset, the Dice of the proposed method reached 0.9727, which is 4.78, 3.81, 1.84, and 2.42% higher than that of 3D UNet, VNet, 3D ResUNet, and 3D DenseUNet, respectively. The 3D UNet showed unsatisfactory results, mainly because the pooling causes the loss of image details during the

downsampling. In addition, the VNet that uses convolution instead of pooling improves the accuracy by 0.97% compared to 3D UNet. Both 3D ResUNet and 3D DenseUNet using residual blocks and densely connected blocks achieve good segmentation results. Moreover, our proposed method achieves superior performance on other metrics except for RVD.

4.3.2. Qualitative comparison results on the LiTS17 dataset

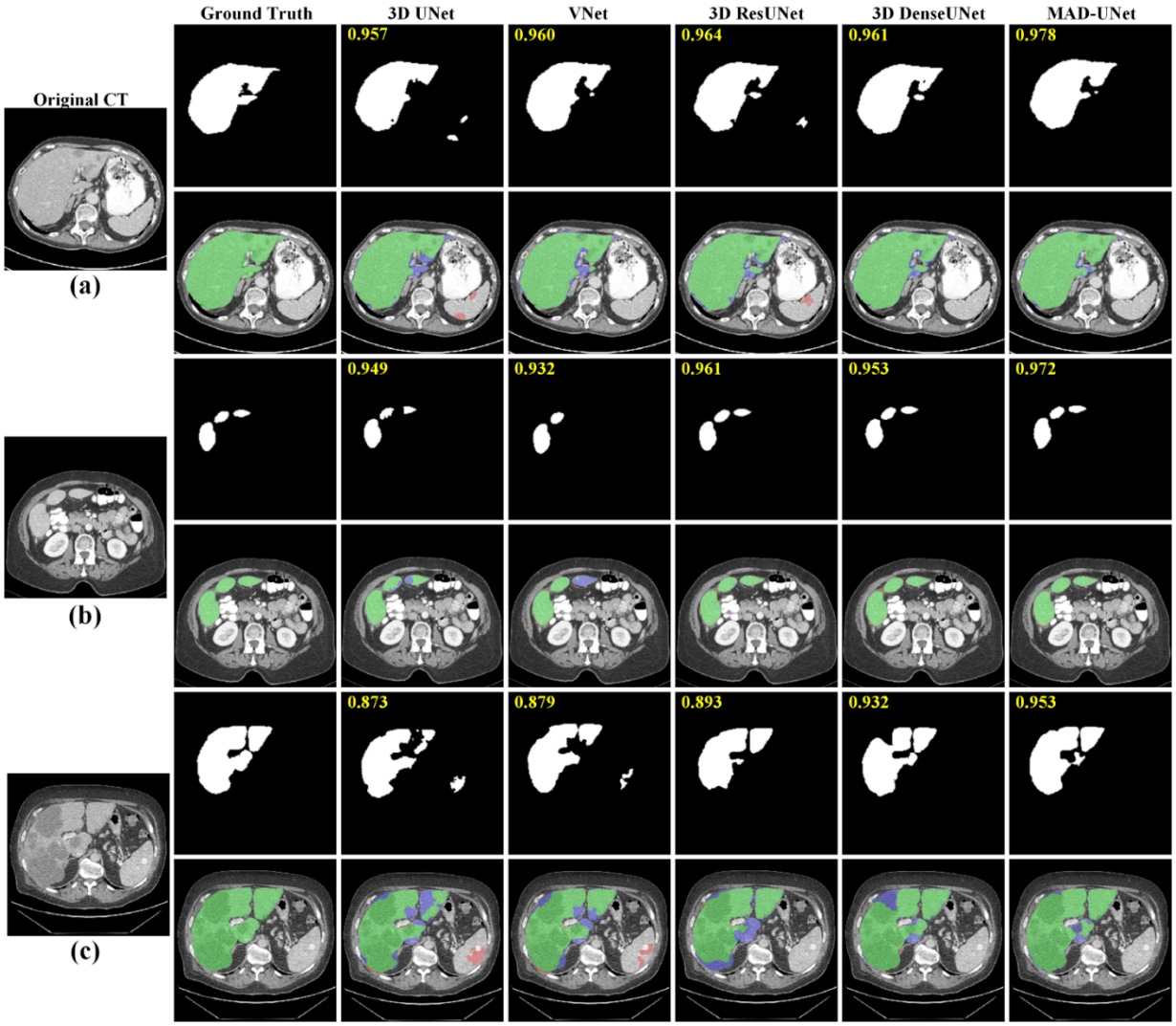

Figure 5. 2D Visualization results of five models on the LiTS17 dataset (a) Liver adjacent to other tissues (b) Discontinuous liver regions (c) Liver containing tumors around (green represents ground truth, blue/red represents under-/over- segmentation).

Figure 5 shows the segmentation results of the proposed method and other SOTA methods. (i) When segmenting liver regions with adjacent tissues (Figure 5(a)), 3D UNet, VNet, and 3D ResUNet mistakenly segmented the spleen as the liver. (ii) When dealing with discontinuous liver regions (Figure 5(b)), 3D UNet and VNet showed results in under-segmentation errors, while MAD-UNet, 3D ResUNet, and 3D DenseUNet accurately segmented the liver. (iii) For the liver region containing the tumor around (Figure 5(c)), the other four networks all produced significant segmentation errors. However, our proposed MAD-UNet showed a slight under-segmentation error.

4.4. Test on the SLiver07 dataset

4.4.1. Quantitative comparison on the SLiver07 dataset

Figure 6 shows the Dice and Loss curves during training on the SLiver07 dataset. Empirically we performed the learning rate decay at epoch 400, and the Dice curve of our proposed MAD-UNet showed a significant drop. However, the Dice value gradually recovered stable and was higher than other models with the epoch increase. The Loss of MAD-UNet also fluctuated in the first 400 epochs of training and then gradually stabilized and was lower than other models.

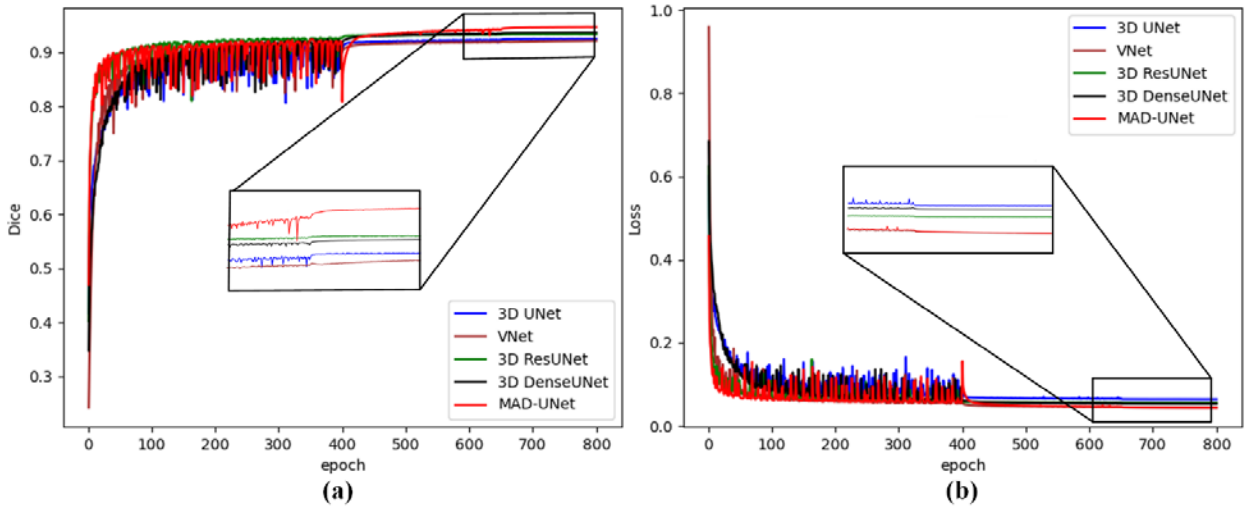

Figure 6. Dice and Loss of different models during training on SLiver07 dataset (a) Dice (b) Loss.

As shown in Table 4, on the SLiver07 dataset, our method achieves the best scores on all five-evaluation metrics. For example, the Dice of MAD-UNet is 4.72, 2.44, 1.5, and 1.19% higher than that of the other four methods, respectively. Besides, our proposed MAD-UNet also showed superior performance on the other four evaluation metrics and thus proved high accuracy and good robustness.

Table 4. Comparative results with SOTA methods on the SLiver07 dataset.

Method	Dice (%)	VOE (%)	RVD (%)	ASD (mm)	RMSD (mm)
3D UNet [17]	92.80 ± 4.13*	13.20 ± 2.85	1.26 ± 0.45	9.93 ± 5.86*	18.51 ± 14.83
VNet [18]	95.08 ± 4.33*	9.12 ± 7.41	-0.93 ± 0.32	3.28 ± 5.12*	7.67 ± 11.06
3D ResUNet [36]	96.02 ± 2.54*	7.55 ± 4.53	-0.66 ± 0.19	2.84 ± 2.56*	9.33 ± 9.43
3D DenseUNet [24]	96.33 ± 1.52*	7.03 ± 2.81	0.58 ± 0.16	3.50 ± 2.67*	12.55 ± 10.79
MAD-UNet	**97.52 ± 0.81**	**4.97 ± 1.73**	**0.23 ± 0.17**	**1.13 ± 0.82**	**4.73 ± 5.21**

Note: Bold font represents the best results. * indicates a statistically significant difference between the labeled results and the corresponding results of our method at a significance level of 0.05.

4.4.2. Qualitative comparison on the SLiver07 dataset

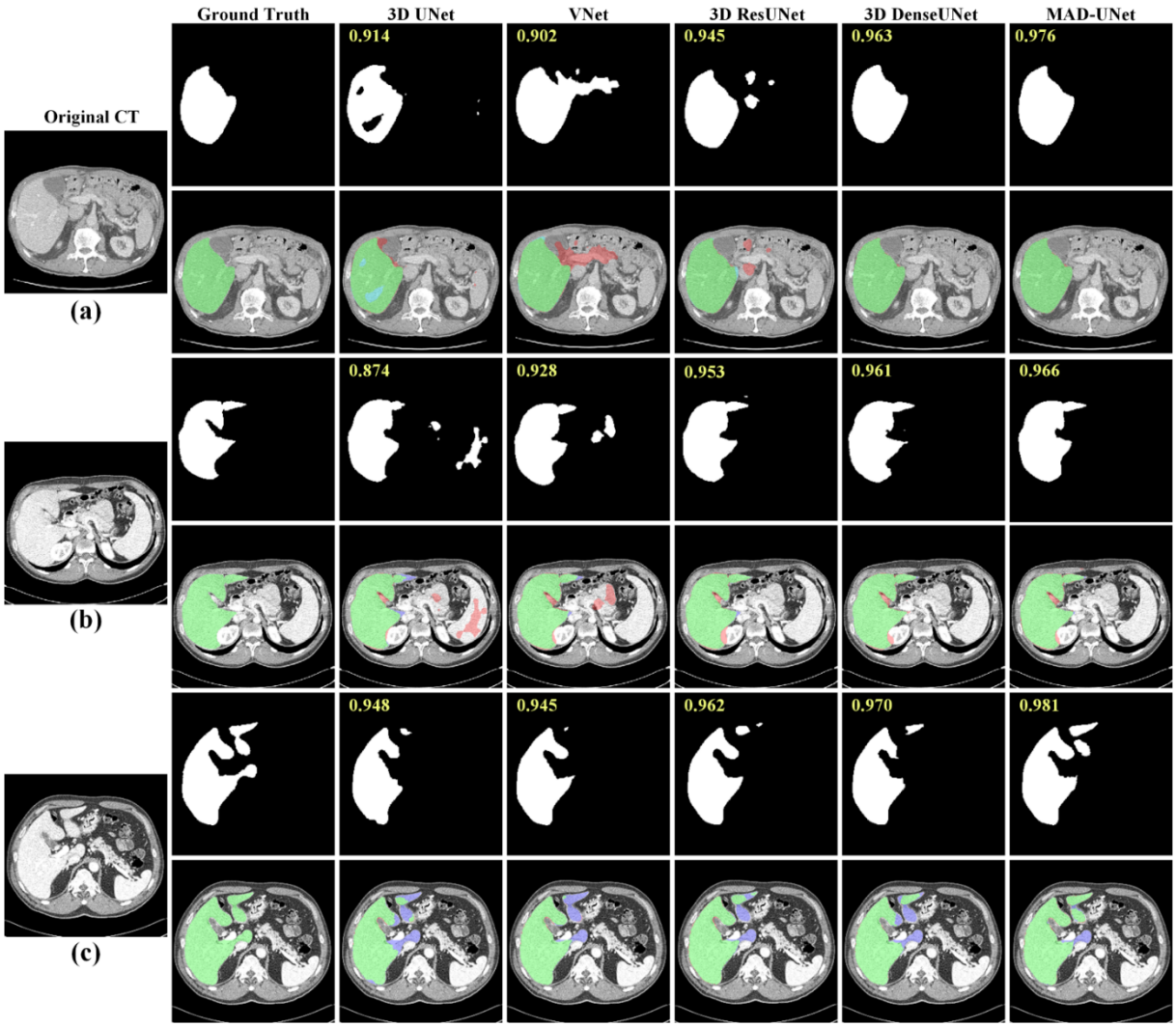

Figure 7. 2D Visualization results of five methods on SLiver07 dataset: (a) Liver containing adjacent tissues; (b) Liver CT intensity similar to that of other organs; (c) Liver with discontinuous regions (green represents ground truth, blue/red represents under-/over- segmentation).

Figure 7 provides the visual segmentation results of different methods on the SLiver07 dataset. (i) When the liver region containing adjacent tissues (Figure 7(a)), 3D UNet, VNet, and 3D ResUNet showed obvious over-segmentation errors; on the contrary, our proposed method and 3D DenseUNet obtained comparable results to the ground truth. (ii) When dealing with the liver CT intensity similar to that of other organs (e.g., spleen, Figure7(b)), both 3D UNet and VNet showed significant over-segmentation errors. However, our method achieves the highest segmentation accuracy. (iii) For liver with discontinuous regions (Figure7(c)), MAD-UNet results in the slightest segmentation error compared with other methods.

4.5. Test results on the 3DIRCADb dataset

4.5.1. Quantitative comparison on the 3DIRCADb dataset

Figure 8 provides the Dice and Loss curves for different models during training on the 3DIRCADb dataset. It can be seen that the Dice of MAD-UNet converges to the highest. Besides, its Loss is always located at the lowest position among all the models.

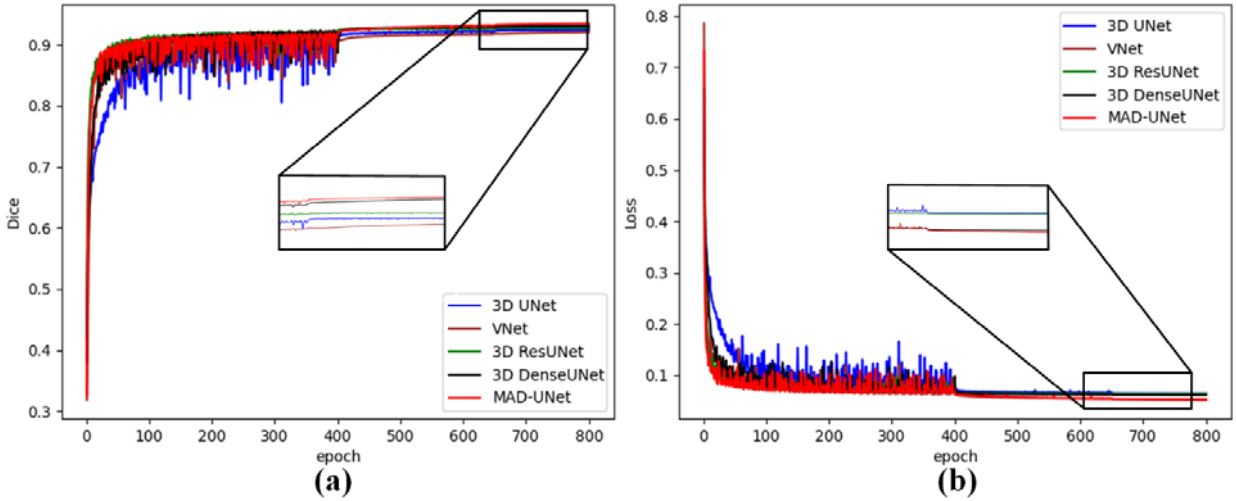

Figure 8. Dice and Loss of different models during training on 3DIRCADb dataset (a) Dice (b) Loss.

Table 5. Comparative results of different methods on the 3DIRCADb dataset.

Method	Dice (%)	VOE (%)	RVD (%)	ASD (mm)	RMSD (mm)
3D UNet [17]	89.88 ± 11.42*	16.95 ± 15.82	2.63 ± 0.98	7.21 ± 6.74*	10.24 ± 9.27
VNet [18]	92.47 ± 5.34*	13.62 ± 8.67	1.05 ± 0.68	4.06 ± 3.30*	7.41 ± 6.43
3D ResUNet [36]	94.61 ± 2.08*	10.16 ± 3.73	**-0.18 ± 0.29**	2.27 ± 1.08*	4.93 ± 3.87
3D DenseUNet [24]	94.56 ± 2.23*	10.25 ± 3.99	1.23 ± 0.75	3.23 ± 1.96*	5.81 ± 4.63
MAD-UNet	**96.91 ± 0.68**	**5.64 ± 1.96**	0.25 ± 0.43	**1.08 ± 0.77**	**2.33 ± 1.15**

*Note: Bold font represents the best results. * indicates a statistically significant difference between the labeled results and the corresponding results of our method at a significance level of 0.05.

As shown in Table 5, on the 3DIRCADb dataset, the effect of 3D UNet results in the worst Dice score (0.8988), while our proposed method obtains Dice of 0.9691. Moreover, our proposed MAD-UNet achieves the best segmentation results on other metrics except for slight inferiority to 3D ResUNet on RVD.

4.5.2. Qualitative comparison on the 3DIRCADb dataset

Figure 9 demonstrates some typical results on the 3DIRCADb dataset. (i) Figure 9(a) presents a segmentation comparison when dealing with discontinuous liver regions that contain tumors at the edges. It can be seen that all five methods result in some over-segmentation in the fuzzy connecting

area of the liver regions. Still, the MAD-UNet showed a relatively more minor error (ii) Figure 9(b) and Figure 9(c) give the segmentation comparison when the liver region contains adjacent organs or tissues. It can be seen that MAD-UNet and 3D DenseUNet achieved relatively stable and high-precision segmentation accuracy.

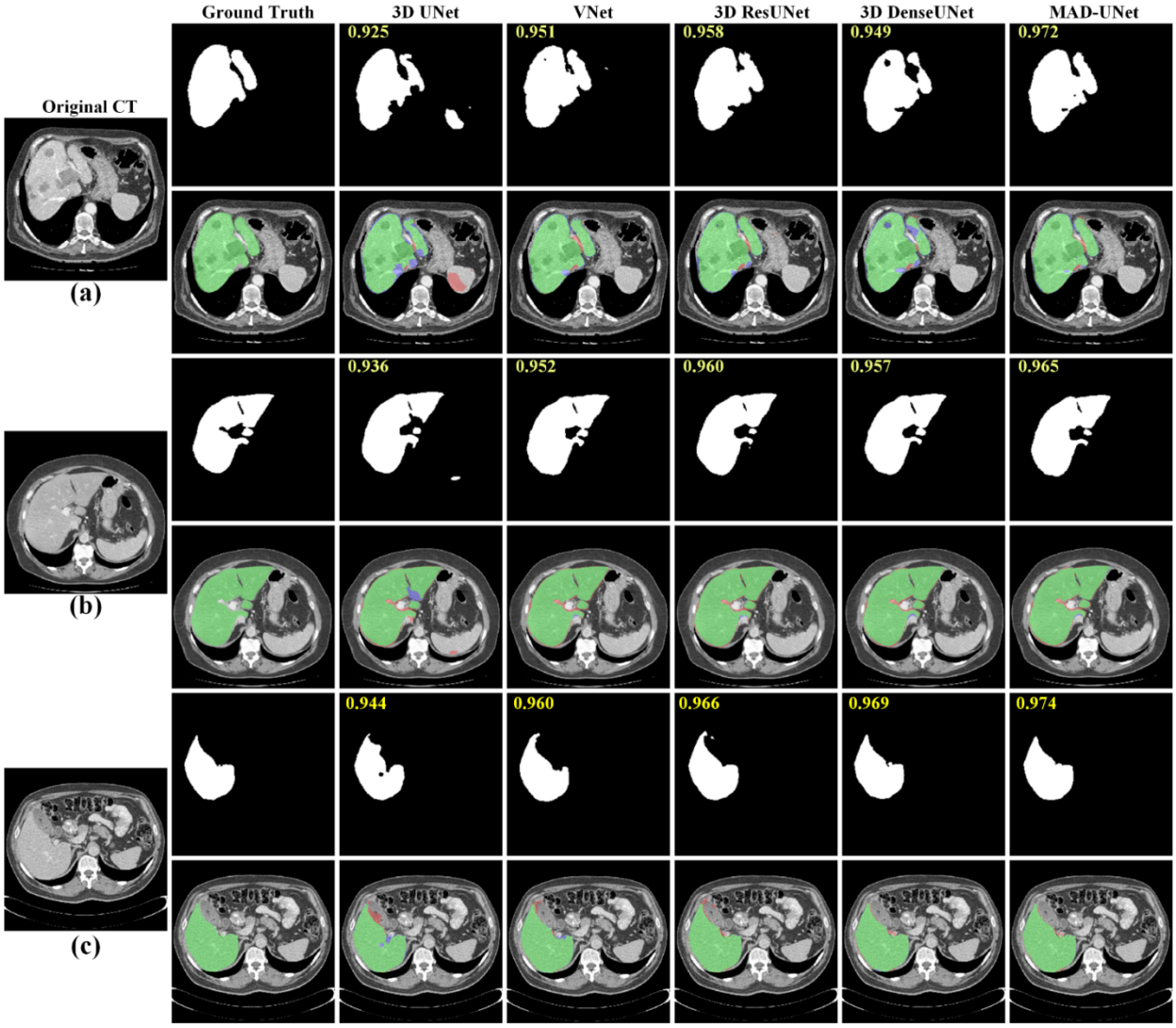

Figure 9. 2D Visual results of different methods on the 3DIRCADb (a) Discontinuous liver region containing tumor at the edge (b) Liver containing adjacent tissue (c) Liver edge containing adjacent organs (green represents ground truth, blue/red represents under-/over- segmentation error).

4.6. Comparison of complexity with other SOTA methods

Table 6 lists the parameters, training, and test time of different methods on LiTS17, SLiver07, and 3DIRCADb datasets. As can be seen from the table, 3D UNet requires the smallest amount of parameters, VNet requires the largest, while deploying the proposed MAD-UNet needs slightly more parameters than 3D UNet and 3D ResUNet. Furthermore, on all three datasets, the training

time of 3D UNet is the least, while that of 3D DenseUNet is the most. Specifically, our proposed method requires the least time to test the three datasets.

Table 6. Comparison of the complexity of different methods on three datasets.

Method	Parameters	LiTS17		SLiver07		3DIRCADb	
		Train time	Test time	Train time	Test time	Train time	Test time
3D UNet [17]	6,405,827	**79h58m4s**	51.54 s	**25h21m12s**	20.42 s	**17h56m51s**	16.71 s
VNet [18]	53,782,217	119h50m21s	52.46 s	38h4m52s	19.63 s	26h48m38s	15.74 s
3D ResUNet [36]	9,498,195	81h24m18m	52.01 s	25h46m45s	18.75 s	18h7m53s	14.42 s
3D DenseUNet [24]	19,783,361	150h19m15s	58.21 s	47h22m43s	19.39 s	33h27m43m	17.89 s
MAD-UNet	9,990,960	85h54m59s	**49.68 s**	27h16m23s	**18.55 s**	19h31m51s	**13.72 s**

4.7. Limitation

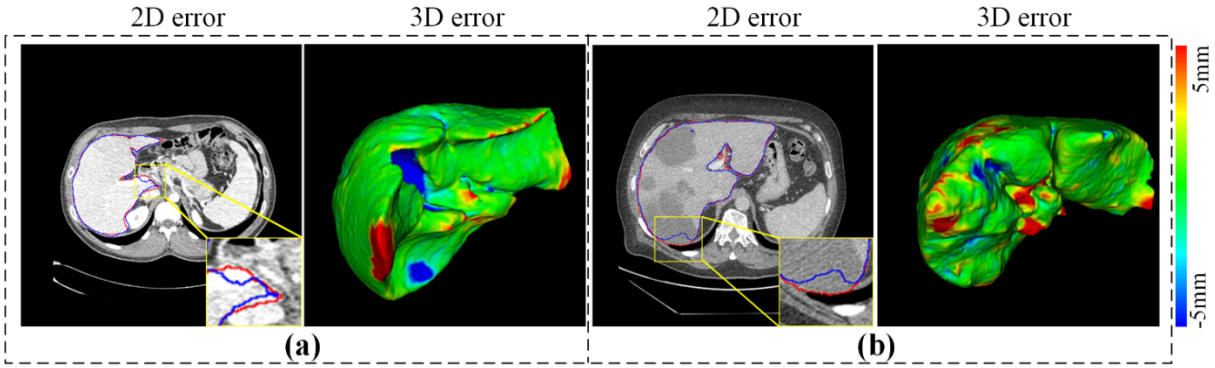

Figure 10. Illustrations of some limitations of proposed MAD-UNet. (a) Liver with blood vessels around (b) Liver with large tumors around (For 2D error, red/blue represents the ground truth and our results, respectively. For 3D error, green represents the ground truth, and blue/red indicates the under-/over- segmentation error).

To illustrate the proposed method's limitation, we present the visualizations of 2D and 3D segmentation errors in some typical cases of the proposed MAD-UNet in Figure 10. (i) From Figure 10(a), we can see that MAD-UNet showed obvious over-segmentation errors in liver regions containing blood vessels. The main reason is that the boundary between the liver and surrounding organs and tissues is blurred in this case. (ii) As shown in Figure 10(b), MAD-UNet produces obvious over-segmentation errors when processing liver regions with large tumors at the edges. This is because the grayscale difference between the tumor and the liver may cause the tumor located at the edge of the liver to be considered as other organs or tissues, resulting in segmentation errors. To alleviate this limitation, we would focus more on strategies to address blurred boundaries, e.g., Zhang et al. [37] effectively suppressed the inconsistency of data distribution by removing mean energy in the preprocessing stage.

5. Conclusions

This paper proposes a new framework by aggregating multi-scale attention and combining it with deep supervision. We aim to improve the liver segmentation accuracy from CT by the proposal of (i) the residual and skip connections, which avoid the repetition of low-resolution feature information and can effectively preserve the edge information of the target. (ii) the attention module, which fully aggregates feature information of different scales and dimensions. It guides the refined attention module to filter out noisy areas, helping the network to pay more attention to areas of interest. (iii) the deep supervision signals, which are used to improve the network's learning ability at different levels.

We extensively validated the proposed method on three publicly available datasets. The experimental results demonstrated that: 1) Compared with the existing SOTA models, our method achieves the best results in four evaluation metrics (Dice, VOE, ASD, and RMSD), with the least test time. 2) The proposed MAD-UNet obtains more satisfactory segmentation performance in dealing with cases, including (i) discontinuous liver regions and (ii) livers containing adjacent tissues or organs.

Nevertheless, the proposed method still shows certain limitations when processing the liver with blood vessels around or containing large tumors at the liver edge. Moreover, the 3D mode requires large memory. Therefore, we will focus on optimizing the network composition to improve the accuracy and robustness of the proposed method via more effective learning of edge features information.

Acknowledgments

This work is supported by the National Nature Science Foundation (No. 61741106, 61701178). This article does not contain any studies with live human participants or animals performed by any of the authors.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. H. A. Nugroho, D. Ihtatho, H. Nugroho, Contrast enhancement for liver tumor identification, in *MICCAI Workshop*, **41** (2008), 201. https://doi.org/10.54294/1uhwld
2. D. Wong, J. Liu, Y. Fengshou, Q. Tian, W. Xiong, J. Zhou, et al., A semi-automated method for liver tumor segmentation based on 2D region growing with knowledge-based constraints, *MICCAI Workshop*, **41** (2008), 159. https://doi.org/10.54294/25etax
3. Y. Yuan, Y. Chen, C. Dong, H. Yu, Z. Zhu, Hybrid method combining superpixel, random walk and active contour model for fast and accurate liver segmentation, *Comput. Med. Imaging Graphics*, **70** (2018), 119–134. https://doi.org/10.1016/j.compmedimag.2018.08.012

4. J. Wang, Y. Cheng, C. Guo, Y. Wang, S. Tamura, Shape-intensity prior level set combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images, *Int. J. Comput. Assisted Radiol. Surg.*, **11** (2016), 817–826. https://doi.org/10.1007/s11548-015-1332-9
5. C. Shi, M. Xian, X, Zhou, H. Wang, H. Cheng, Multi-slice low-rank tensor decomposition based multi-atlas segmentation: Application to automatic pathological liver CT segmentation, *Med. Image Anal.*, **73** (2021), 102152. https://doi.org/10.1016/j.media.2021.102152
6. Z. Yan, S. Zhang, C. Tan, H. Qin, B. Belaroussi, H. J. Yu, et al. Atlas-based liver segmentation and hepatic fat-fraction assessment for clinical trials, *Comput. Med. Imaging Graphics*, **41** (2015), 80–92. https://doi.org/10.1016/j.compmedimag.2014.05.012
7. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2015), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683
8. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.
9. O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Cham, (2015), 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
10. Y. Liu, N. Qi, Q. Zhu, W. Li, CR-U-Net: Cascaded U-net with residual mapping for liver segmentation in CT images, in *IEEE Visual Communications and Image Processing (VCIP)*, (2019), 1–4. https://doi.org/10.1109/VCIP47243.2019.8966072
11. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778.
12. X. Xi, L. Wang, V. Sheng, Z. Cui, B. Fu, F. Hu, Cascade U-ResNets for simultaneous liver and lesion segmentation, *IEEE Access*, **8** (2020), 68944–68952. https://doi.org/10.1109/ACCESS.2020.2985671
13. O. Oktay, J. Schlemper, L. Folgoc, M. Lee, M. Heinrich, K. Misawa, et al., Attention u-net: Learning where to look for the pancreas, preprint, arXiv:1804.03999.
14. L. Hong, R. Wang, T. Lei, X. Du, Y. Wan, Qau-Net: Quartet attention U-net for liver and liver-tumor segmentation, in *IEEE International Conference on Multimedia and Expo (ICME)*, (2021), 1–6. https://doi.org/10.1109/ICME51207.2021.9428427
15. W. Cao, P. Yu, G. Lui, K. W. Chiu, H. M. Cheng, Y. Fang, et al., Dual-attention enhanced BDense-UNet for liver lesion segmentation, preprint, arXiv:2107.11645.
16. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, **35** (2012), 221–231. https://doi.org/10.1109/TPAMI.2012.59
17. Ö. Çiçek, A. Abdulkadir, S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: Learning dense volumetric segmentation from sparse annotation, in *International Conference on Medical Image Computing and Computer-assisted Intervention*, Springer, Cham, (2016), 424–432. https://doi.org/10.1007/978-3-319-46723-8_49
18. F. Milletari, N. Navab, S. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, in *International Conference on 3D Vision (3DV)*, (2016), 565–571. https://doi.org/10.1109/3DV.2016.79

19. Z. Liu, Y. Song, V. Sheng, L. Wang, R. Jiang, X. Zhang, et al., Liver CT sequence segmentation based with improved U-Net and graph cut, *Expert Syst. Appl.*, **126** (2019), 54–63. https://doi.org/10.1016/j.eswa.2019.01.055
20. T. Lei, W. Zhou, Y. Zhang, R. Wang, H. Meng, A. Nandi, Lightweight v-net for liver segmentation, in *ICASSP 2020—2020 IEEE International Conference on Acoustics*, *Speech and Signal Processing (ICASSP)*, (2020), 1379–1383. https://doi.org/10.1109/ICASSP40776.2020.9053454
21. T. Zhou, L. Li, G. Bredell, J. Li, E. Konukoglu, Volumetric memory network for interactive medical image segmentation, *Med. Image Anal.*, **2022** (2022), 102599. https://doi.org/10.1016/j.media.2022.102599
22. Q. Jin, Z. Meng, C. Sun, H. Cui, R. Su, RA-UNet: A hybrid deep attention-aware network to extract liver and tumor in CT scans, *Front. Bioeng. Biotechnol.*, **2020** (2020), 1471. https://doi.org/10.3389/fbioe.2020.605132
23. X. Han, Automatic liver lesion segmentation using a deep convolutional neural network method, preprint, arXiv:1704.07239.
24. X. Li, H. Chen, X. Qi, Q. Dou, C. Fu, P. Heng, H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes, *IEEE Trans. Med. Imaging*, **37** (2018), 2663–2674. https://doi.org/10.1109/TMI.2018.2845918
25. P. Lv, J. Wang, H. Wang, 2.5D lightweight RIU-Net for automatic liver and tumor segmentation from CT, *Biomed. Signal Process. Control*, **75** (2022), 103567. https://doi.org/10.1016/j.bspc.2022.103567
26. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 7132–7141.
27. S. Woo, J. Park, J. Lee, I. Kweon, Cbam: Convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
28. W. Li, Y. Tang, Z. Wang, K. Yu, S. To, Atrous residual interconnected encoder to attention decoder framework for vertebrae segmentation via 3D volumetric CT images, *Eng. Appl. Artif. Intell.*, **114** (2022), 105102. https://doi.org/10.1016/j.engappai.2022.105102
29. T. Zhou, J. Li, S. Wang, R. Tao, J. Shen, Matnet: Motion-attentive transition network for zero-shot video object segmentation, *IEEE Trans. Image Process.*, **29** (2020), 8326–8338. https://doi.org/10.1109/TIP.2020.3013162
30. Y. Wang, H. Dou, X. Hu, L. Zhu, X. Yang, M. Xu, et al., Deep attentive features for prostate segmentation in 3D transrectal ultrasound, *IEEE Trans. Med. Imaging*, **38** (2019), 2768–2778. https://doi.org/10.1109/TMI.2019.2913184
31. C. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, PMLR, (2015), 562–570.
32. Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, et al., 3D deeply supervised network for automated segmentation of volumetric medical images, *Med. Image Anal.*, **41** (2017), 40–54. https://doi.org/10.1016/j.media.2017.05.001
33. B. Wang, Y. Lei, S. Tian, T. Wang, Y. Liu, P. Patel, et al., Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation, *Med. Phys.*, **46** (2019), 1707–1718. https://doi.org/10.1002/mp.13416

34. J. Yang, B. Wu, L. Li, P. Cao, O. Zaiane, MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT, *Comput. Med. Imaging Graphics*, **92** (2021), 101957. https://doi.org/10.1016/j.compmedimag.2021.101957
35. T. Heimann, B. Van Ginneken, M. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, et al. Comparison and evaluation of methods for liver segmentation from CT datasets, *IEEE Trans. Med. Imaging*, **28** (2009), 1251–1265. https://doi.org/10.1109/TMI.2009.2013851
36. W. Xu, H. Liu, X. Wang, Y. Qian, Liver segmentation in CT based on ResUNet with 3D probabilistic and geometric post process, in *IEEE 4th International Conference on Signal and Image Processing (ICSIP)*, (2019), 685–689. https://doi.org/10.1109/SIPROCESS.2019.8868690
37. C. Zhang, Q. Hua, Y. Chu, P. Wang, Liver tumor segmentation using 2.5D UV-Net with multi-scale convolution, *Comput. Biol. Med.*, **133** (2021), 104424. https://doi.org/10.1016/j.compbiomed.2021.104424

AIMS Press