Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Iterative Convolutional Encoder-Decoder Network with Multi-Scale Context Learning for Liver Segmentation

Feiyan Zhang, Shuhao Yan, Yizhong Zhao, Yuan Gao, Zhi Li & Xuesong Lu

To cite this article: Feiyan Zhang, Shuhao Yan, Yizhong Zhao, Yuan Gao, Zhi Li & Xuesong Lu (2022) Iterative Convolutional Encoder-Decoder Network with Multi-Scale Context Learning for Liver Segmentation, Applied Artificial Intelligence, 36:1, 2151186, DOI: 10.1080/08839514.2022.2151186

To link to this article: https://doi.org/10.1080/08839514.2022.2151186

Published online: 08 Dec 2022.

APPLIED ARTIFICIAL INTELLIGENCE
2022, VOL. 36, NO. 1, e2151186 (3943 pages)
https://doi.org/10.1080/08839514.2022.2151186

Taylor & Francis
Taylor & Francis Group

Iterative Convolutional Encoder-Decoder Network with Multi-Scale Context Learning for Liver Segmentation

Feiyan Zhang[a], Shuhao Yan[b], Yizhong Zhao[a], Yuan Gao[a], Zhi Li[a], and Xuesong Lu[a]

[a]College of Biomedical Engineering, South-Central Minzu University, Wuhan, P. R. China; [b]Xiangyang Central Hospital, Affiliated Hospital of Hubei University of Arts and Science, Xiangyang, P. R. China

ABSTRACT

Rapid and accurate extraction of liver tissue from abdominal computed tomography (CT) and magnetic resonance (MR) images has critical importance for diagnosis and treatment of hepatic diseases. Due to adjacent organs with similar intensities and anatomical variations between different subjects, the performance of segmentation approaches based on deep learning still has room for improvement. In this study, a novel convolutional encoder-decoder network incorporating multi-scale context information is proposed. The probabilistic map from previous classifier is iteratively fed into the encoder layers, which fuses high-level shape context with low-level appearance features in a multi-scale manner. The dense connectivity is adopted to aggregate feature maps of varying scales from the encoder and decoder. We evaluated the proposed method with 2D and 3D application on abdominal CT and MR images of three public datasets. The proposed method generated liver segmentation with significantly higher accuracy (p <0.05), in comparison to several competing methods. These promising results suggest that the novel model could offer high potential for clinical workflow.

ARTICLE HISTORY
Received 15 September 2022
Revised 9 November 2022
Accepted 18 November 2022

Introduction

Liver segmentation on medical images plays a critical role in hepatic disease diagnosis, function assessment, radiotherapy planning, and image-guided surgery. In clinical workflow, computed tomography (CT) is the most common technique for detecting numerous types of malignant liver tumors (Chen et al. 2011). On the other hand, due to non-ionizing radiation and better contrast of soft tissues, magnetic resonance (MR) imaging is increasingly used to monitor liver volume and fat content, which could aid in reducing the need of more invasive biopsies (Tang et al. 2015).

Manual delineation of the liver is time-consuming and prone to incur inter-observer variations. Semi-automatic or automatic approaches have been developed for radiologists and physicians (Chartrand et al. 2017; Moghbel et al. 2018). However, it is still a challenging task to rapidly and accurately

CONTACT Xuesong Lu 365103248@qq.com College of Biomedical Engineering, South-Central Minzu University, Wuhan 430074, P. R. China

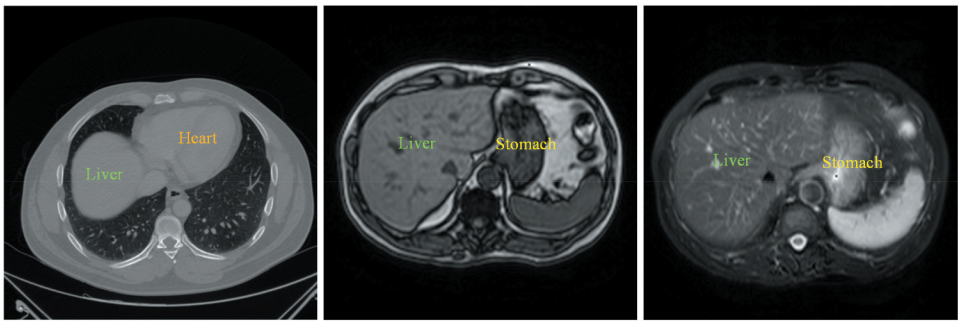

Figure 1. Examples of CT and MR abdominal organ images. (a) a CT image slice in axial view. (b) a T1-weighted sequence of MR image slice in axial view. (c) a T2-weighted sequence of MR image slice in axial view.

extract liver organ from both CT and MR images. As can be seen in Figure 1, several adjacent organs such as heart and stomach share similar intensities with liver. The multiple sequences of MR often suffer from image artifacts and intensity inhomogeneities. Moreover, the shape, size, and texture of liver in both CT and MR vary a lot across different subjects.

In the last few decades, most studies in liver segmentation have mainly focused on five types of methods (Heimann et al. 2009): statistical shape model, level set, graph cut, multi-atlas label fusion, and machine learning. Cerrolaza *et al.* introduced a generalized multi-resolution point distribution model to describe abdominal multi-organ shape, which can be integrated into the active shape-model-based segmentation (Cerrolaza et al. 2015). Wang *et al.* incorporated a shape-intensity prior model into level set framework for liver segmentation (Wang et al. 2016). Li *et al.* proposed the deformable graph cut based on shape constraints to accurately detect the

liver surface (Li et al. 2015). Chen *et al.* integrated the shape prior from multi-atlas label fusion with graph cut technique to reduce the errors of segmentation on abdominal CT images (Chen et al. 2020). Jin *et al.* applied generalized Hough transform and active appearance model to localize the renal cortex, and then a modified random forests method was employed to segment the kidney into four components (Jin et al. 2016).

Recently, deep learning methods, in particular convolutional neural network (CNN), have been used successfully in medical image segmentation (Ker et al. 2018; Litjens et al. 2017; Shen, Wu, and Suk 2017). A fully convolutional network (FCN) (Long et al. 2015) was trained end-to-end, pixels-to-pixels on semantic segmentation. In order to yield precise segmentation, a U-shaped architecture network (U-Net) (Ronneberger et al. 2015) was designed to combine encoding and decoding feature maps with skip connections. A practical deep convolutional encoder-decoder network for image segmentation (SegNet) (Badrinarayanan, Kendall, and Cipolla 2017) was presented

using nonlinear upsampling with pooling indices. A dense V-network (DenseVNet) (Gibson et al. 2018) that enables high-resolution activation maps through memory-efficient dropout and feature reuse was proposed for abdominal multi-organ segmentation. To overcome the restrictive feature fusion scheme in U-Net, a novel architecture so-called UNet++ (Zhou et al. 2020) redesigning skip connections was presented for accurate image segmentation, which introduces a build-in ensemble of U-Nets of varying depths.

According to human perception, when someone observes a scene, the eyes move along the whole visual space, and then concentrate on region of interest (Zhang et al. 2017). Context information (Chen et al. 2016) mimicking this property of human perception has shown to be useful in image segmentation. Salehi *et al.* implemented an auto-context convolutional neural network upon the U-Net architecture for brain extraction. The posterior probability maps from the network output were utilized iteratively as context information to learn the local shape (Salehi, Erdogmus, and Gholipour 2017). Zhang *et al.* efficiently combined features within a single CT image and among multiple adjacent images for multi-organ segmentation (Zhang et al. 2018). Oktay *et al.* proposed a novel attention gate model that automatically learns to focus on target structures of varying shapes and sizes for medical imaging (Oktay et al. 2018). Yang *et al.* exploited the bidirectional long-short term memory network (BiLSTM) that can capture contextual cues to refine ultrasound segmentation (Yang et al. 2019). Since the intrinsic locality of convolution operations, Chen *et al.* integrated Transformers into U-Net framework as a strong alternative for medical image segmentation (Chen et al. 2022). Cao *et al.* developed a U-shaped pure Transformer (Swin-Unet) for multi-organ segmentation of abdominal and cardiac images (Cao et al. 2022). Hatamizadeh *et al.* introduced a 3D Transformer (UNETR) as the encoder to learn sequence representations of the input volume (Hatamizadeh et al. 2022).

To keep the model relatively simple, in this study we develop a novel convolutional encoder-decoder network incorporating multi-scale context information and apply it to liver segmentation. The probabilistic map from output layer of decoder part is iteratively fed into the encoder layers, which fuses high-level shape and context information with low-level appearance features in a multi-scale manner. Furthermore, the dense connectivity like UNet++ is adopted to aggregate feature maps of varying scales from the encoder and decoder. The proposed method for liver segmentation is evaluated on abdominal CT with 2D application and abdominal MR with 3D application.

Methods

Network Architecture

For the challenging liver segmentation of abdominal images, the architecture of classical U-Net with the pooling layer is prone to lose the information of

image details in the downsampling step. In order to improve network performance, the posterior probabilities from the previous classifier are considered as features, and are merged into the proposed network. As a result, the substructure information would be compensated for the downsampling procedures in iterative and multi-scale manner.

As shown in Figure 2, the proposed network consists of convolutional encoder part and decoder part. The basic unit of node $X^{i,j}$ where i indexes the downsampling layer along the encoder and j indexes the convolution layer along the skip connection is the convolutional block. For 2D application, the convolutional block is composed of two consecutive padded 3×3 convolutions followed by ReLU (Rectified Linear Unit) (Nair and Hinton 2010) layers. At node $X^{0,0}$, the original image x and the probabilistic map S_{t-1} are input to the convolutional

block, respectively. After that the feature maps from these blocks are concatenated, a 2×2 max-pooling operation with stride 2 is applied. In order to prepare for next scale connection, the S^{ℓ}_{t-1} from the posterior probabilities are transferred to a max-pooling layer, which are used to node $X^{1,0}$. Therefore the probabilistic map in two scales ($\ell = 1$) is fed into the network for guiding the segmentation task. After each downsampling operation, the number of feature channels is doubled.

In the decoder part, a 2×2 transpose convolutional layer (Dumoulin and Visin 2018) for upsampling operation is applied after the convolutional block. Inspired by UNet++, the U-Nets of varying depths are realized through the extended decoders and unified into the ensemble architecture. With dense connectivity, each decoder fuses the final aggregated feature maps and the intermediate aggregated feature maps, as well as the same-scale feature maps from the encoder. As such, the multi-scale context information can be propagated to the aggregation layers across the network. After each upsampling operation, the number of feature channels is halved. Without deep supervision, a 1×1 convolution is appended

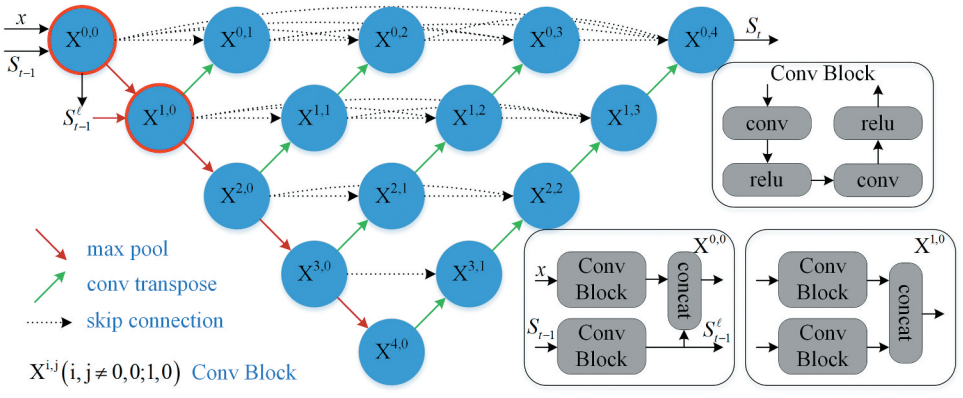

Figure 2. Network architecture.

to the final layer. For 3D application, the kernel sizes of convolution and max-pooling layers are correspondingly extended to $3 \times 3 \times 3$ and $2 \times 2 \times 2$.

Technical Details

In the proposed segmentation method, a sequence of classifiers ($t = 1, \ldots, K$) is designed in an iterative way, in which the posterior probabilities from the previous classifier are incorporated into the current process as features. Let $\{x_i, l_i, i = 1\}$ be m training image pairs, where x_i and l_i are respectively the intensity image and corresponding label image. For each image at step t, the pair of $\left\{x_i, s_i^{(t-1)}\right\}$ is delivered into the network for classification where $s_i^{(t-1)}$ is the posterior probability of image x_i from previous step. When step $t = 1$, s_i^0 is constructed with uniform distribution (i.e. the probability value of 0.5 for binary classification).

Algorithm 1. The procedure of iterative learning based on the proposed network

Input: $\{x_i, l_i, i = 1\}$ the training image pairs; s_i^0 the initial probability map with uniform distribution
Output: the learned weights of the proposed networks (step $t = 1, \ldots, K$)
repeat
Obtain the training set $\Gamma^t = \left\{\left(x_i, s_i^{(t-1)}\right), l_i, i = 1\right\}$;
Train the proposed network described in Figure 2 using Γ^t;
Calculate the posterior probability s_i^t through Sigmoid activation function for $\{i = 1\}$;
Calculate the loss H_i^t via Equation (1) for $\{i = 1\}$;
Calculate $\Delta \mathrm{H} = 1 \sum_{i=1} \left|\mathrm{H}_i^t - \mathrm{H}_i^{(t-1)}\right|$;
until $\Delta \mathrm{H} < \varepsilon$ (ε is a predefined threshold)

For each image, the posterior probability can be calculated through the network in Figure 2 followed by a Sigmoid activation function. During the optimization, a hybrid loss consisting of pixel-wise cross-entropy and soft dice-coefficeint is minimized. Mathematically, it can be defined as:

$$\mathrm{H}(L, S) = -\omega \cdot \frac{1}{N} \sum_{n=1}^{N} l_n \cdot \log s_n + \left(1 - \frac{2 \sum_{n=1}^{N} l_n s_n}{\sum_{n=1}^{N} l_n + \sum_{n=1}^{N} s_n}\right), \qquad (1)$$

where $l_n \in L$ denotes image labels, $s_n \in S$ denotes predicted probabilities, N denotes the number of pixels, and ω is the balance factor. Note that only cross-entropy loss is employed for 2D application. Algorithm 1 shows the procedure of weights learning for the sequence of classifiers. In the test stage, the learned weights can be successively applied to the first two classifiers (K=2) in the sequence for segmentation.

Experiments

Datasets

Three publicly available datasets covering liver organs were used in this study. The first dataset was from 3Dircadb1 (Bilic et al. 2022), which contains 20 contrast-enhanced CT scans. The image size is $512 \times 512 \times 74 \sim 260$ voxels. The in-plane spacing varies from 0.57 mm to 0.87 mm, and slice thickness ranges from 1 mm to 4 mm. The second dataset was from Sliver07 (Heimann et al. 2009), which contains 20 contrast-enhanced CT scans. The image size is $512 \times 512 \times 64 \sim 394$ voxels. The in-plane spacing varies from 0.58 mm to 0.82 mm, and slice thickness ranges from 1 mm to 3 mm.

The third dataset was provided by the Combined Healthy Abdominal Organ Segmentation challenge in 2019 (CHAOS19) (Kavur et al. 2020), including 20 contrast-enhanced CT scans and 20 multi-sequence MR scans. For the CT data, the image size is $512 \times 512 \times 78 \sim 294$ voxels. The in-plane spacing varies from 0.54 mm to 0.79 mm, and slice thickness ranges from 2.0 mm to 3.2 mm. The MR data includes two different sequences T1-DUAL and T2-SPIR. In total, there are 60 images from T1-DUAL in phase (T1-DUALin), T1-DUAL oppose phase (T1-DUALout), and T2-SPIR for 20 patients. The image size is $256 \times 256 \times 26 \sim 50$ voxels. The in-plane spacing varies from 0.72 mm to 2.03 mm, and slice thickness ranges from 4.4 mm to 8.0 mm. The manual delineation of liver tissue in each image is regarded as the ground truth for validation.

Evaluation Metrics

To quantitatively evaluate the performance of the proposed method, we used four metrics (Heimann et al. 2009): the Dice coefficient (DICE), the Relative absolute volume difference (RAVD), the Average symmetric surface distance (ASSD), and the Maximum symmetric surface distance (MSSD). Assuming that V_A denotes segmentation by the algorithm and V_B denotes segmentation by the ground truth, the DICE and RAVD can be defined as follows:

$$DICE = \frac{2|V_A \cap V_B|}{|V_A| + |V_B|}, \tag{2}$$

$$RAVD = \text{abs}\left(\frac{|V_A|}{|V_B|} - 1\right) \cdot 100, \tag{3}$$

where $|\cdot|$ indicates the number of voxels within the segmentation and $abs(\cdot)$ indicates the absolute value. Assuming that S_A denotes the surface of segmentation by the algorithm and S_B denotes the surface of segmentation by the ground truth, the ASSD and MSSD can be defined as follows:

$$ASSD = \frac{1}{|S_A| + |S_B|}\left(\sum dist(S_A, S_B) + \sum dist(S_B, S_A)\right), \tag{4}$$

$$MSSD = \max\{\max dist(S_A, S_B), \max dist(S_B, S_A)\}, \tag{5}$$

where $dist(S_A, S_B)$ indicates the Euclidean distance of the set of points on S_A to the nearest point on S_B. For the DICE, the larger the value is, the better the segmentation result is. For the other metrics, the smaller the value is, the better the segmentation result is. A value of $p < 0.05$ in two-sided Wilcoxon tests was considered to indicate a statistically significant difference between two methods.

Experimental Setup

For the CT images in each dataset, the networks process 2D axial slices, and then the segmentation results are stacked into 3D volumes. A 5-fold cross validation was performed on 20 cases of each dataset for liver segmentation. For comparison, we use the original 2D U-Net (Ronneberger et al. 2015), 2D Auto-Net (Salehi, Erdogmus, and Gholipour 2017), and UNet++ (Zhou et al. 2020) for 2D tasks (2D UNet++) as baseline methods. To avoid over-fitting, the data was augmented by random rotation (between 0 and 90 degrees), random flipping (on two axes), and random elastic deformation (grid displacements from Gaussian distribution with 2 pixels standard deviation).

For the MR images in the CHAOS19 dataset, the networks were operated in 3D mode. A 5-fold cross validation was performed on 20 cases of each sequence (T1-DUALin, T1-DUALout, and T2-SPIR) for liver segmentation. For comparison, we use the 3D U-Net (Cicek et al. 2016), 3D Auto-Net (Salehi, Erdogmus, and Gholipour 2017), and UNet++ (Zhou et al. 2020) for 3D tasks (3D UNet++) as baseline methods. All MR images were resampled to the spacing of $1.5 \times 1.5 \times 6.0$ mm and cropped to the size of $224 \times 224 \times 48$ voxels. Data augmentation including random rotation

(between 0 and 90 degrees), random flipping (on three axes), and random elastic deformation (grid displacements from Gaussian distribution with 2 voxels standard deviation) was used to alleviate over-fitting problem.

Implementation Details

The proposed method was implemented using PyTorch (Paszke et al. 2017) on a PC with an NVIDIA GeForce RTX 3080Ti GPU. Before training and test, both CT and MR were normalized to zero mean and unit variance. Our model was trained using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 1*e*-4, a batch size of 1, and 16 base filters in the first layer. The number of training epochs per step was 40 and $\varepsilon = 10^{-6}$ for 2D application, while 80

epochs were set to each step and $\varepsilon = 10^{-3}$ for 3D application. In addition, the balance factor in Equation (1) was set to $\omega = 0.5$. Our code is freely available at https://github.com/zfy012/Ite-netpp.

Results and Discussions

Results on Abdominal CT Images

Figure 3 shows the learning curves of the 2D UNet++ and the proposed method for liver segmentation on CT images. It is obvious that our method enables a better optimization than 2D UNet++ for the tasks of three datasets. The DICE results of liver segmentation using four methods are plotted in Figure 4. In 3Dircadb1, the median DICE of the proposed method increases significantly compared to 2D UNet++ from 0.946 to 0.956 ($p = 3.50 \times 10^{-3}$). In Sliver07, the median

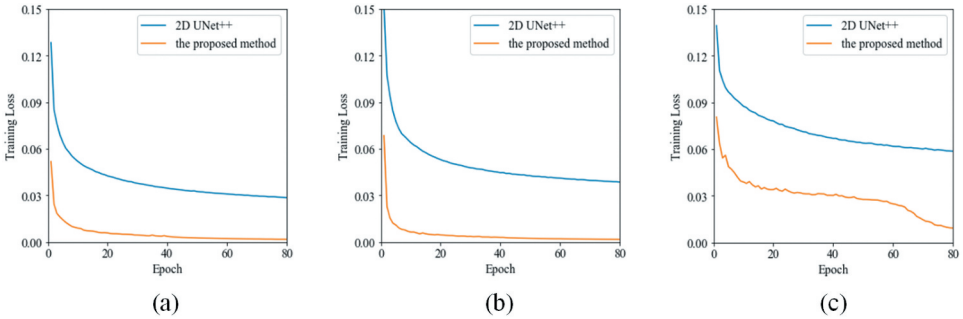

Figure 3. The training results of the 2D UNet++ and the proposed method over 80 epochs from (a) the 3Dircadb1 dataset, (b) the Sliver07 dataset, and (c) the CHAOS19 dataset.

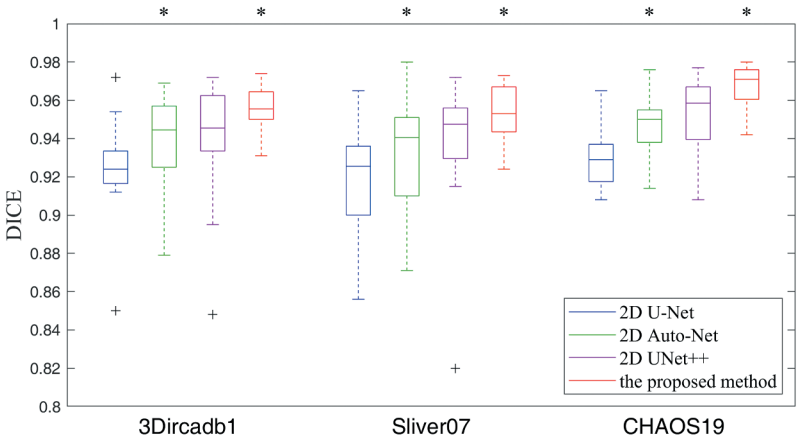

Figure 4. The boxplot of DICE results using four methods on CT images of three datasets. A star indicates a statistical significant difference of the median DICE compared to the previous column.

Table 1. The mean and standard deviation of quantitative measures for liver segmentation on three datasets.

	Methods	DICE	RAVD	ASSD(mm)	MSSD(mm)
3Dircadb1	2D U-Net	0.925±0.023	6.030±3.428	3.927±2.150	55.300±20.047
	2D Auto-Net	0.940±0.021	4.899±3.631	2.997±1.588	**29.457**±6.317
	2D UNet++	0.941±0.028	4.859±5.510	1.277±1.124	38.079±19.871
	the proposed method	**0.955**±0.012	**3.782**±2.306	**1.071**±0.981	32.939±12.252
Sliver07	2D U-Net	0.918±0.032	7.720±5.917	4.839±2.224	55.963±45.400
	2D Auto-Net	0.933±0.031	5.831±4.376	4.183±2.210	39.682±12.647
	2D UNet++	0.938±0.032	4.127±3.369	3.660±2.195	40.635±15.212
	the proposed method	**0.954**±0.014	**3.166**±2.721	**2.059**±0.748	**29.448**±9.285
CHAOS19	2D U-Net	0.931±0.016	6.064±3.460	3.832±1.811	45.633±26.153
	2D Auto-Net	0.947±0.016	4.698±2.437	2.493±1.066	**28.485**±7.028
	2D UNet++	0.952±0.021	3.025±2.231	1.217±0.434	36.245±11.945
	the proposed method	**0.967**±0.011	**2.296**±1.980	**0.929**±0.341	28.948±8.856

DICE of the proposed method increases significantly compared to 2D UNet++ from 0.947 to 0.953 ($p = 2.17 \times 10^{-2}$). In CHAOS19, the median DICE of the proposed method increases significantly compared to 2D UNet ++ from 0.958 to 0.971 ($p = 6.20 \times 10^{-3}$). Table 1 lists the quantitative results of liver segmentation evaluation metrics for three datasets. The proposed method obtained the best DICE, RAVD, and ASSD on all datasets. With regard to MSSD, the 2D Auto-Net achieved the best results on the 3Dircadb1 and CHAOS19 datasets. Figure 5 displays some segmentation results produced by using four methods. It can be seen that the contours through our method are much closer to the liver boundaries of ground truth.

Results on Abdominal MR Images

Figure 6 shows the learning curves of the 3D UNet++ and the proposed method for liver segmentation on multi-sequence MR images of the CHAOS19 dataset. It is clear that our method enables a better optimization than 3D UNet++ for these tasks. The DICE results of liver segmentation using four methods are plotted in Figure 7. In the T1-DUALin sequence, the median DICE of the proposed method increases significantly compared to 3D UNet++ from 0.940 to

0.952 ($p = 2.54 \times 10^{-2}$). In the T1-DUALout sequence, the median DICE of the proposed method increases significantly compared to 3D UNet++ from 0.936 to 0.942 ($p = 4.19 \times 10^{-2}$). In the T2-SPIR sequence, the median DICE of the proposed method increases significantly compared to 3D UNet++ from 0.916 to 0.942 ($p = 2.80 \times 10^{-3}$). Table 2 lists the quantitative results of liver segmentation evaluation metrics for three sequences. The proposed method obtained the best DICE, RAVD, and MSSD on all sequences. With regard to ASSD, the 3D Auto-Net achieved the best results except for the T1-DUALout sequence. Figure 8

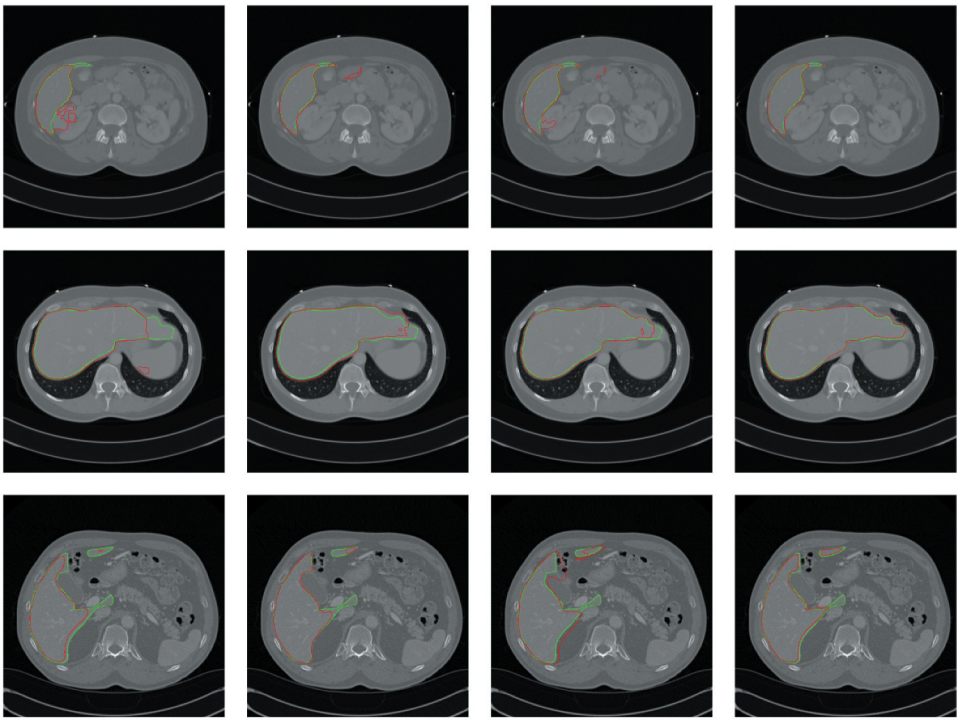

Figure 5. Typical liver segmentation results on CT images of the 3Dircadb1 (top row), the Sliver07 (middle row), and the CHAOS19 (bottom row) datasets by using four methods. (a) 2D U-Net. (b) 2D Auto-Net. (c) 2D UNet++. (d) the proposed method. Green contours indicate the ground truth segmentation, and red contours indicate the automatic segmentation by the algorithm.

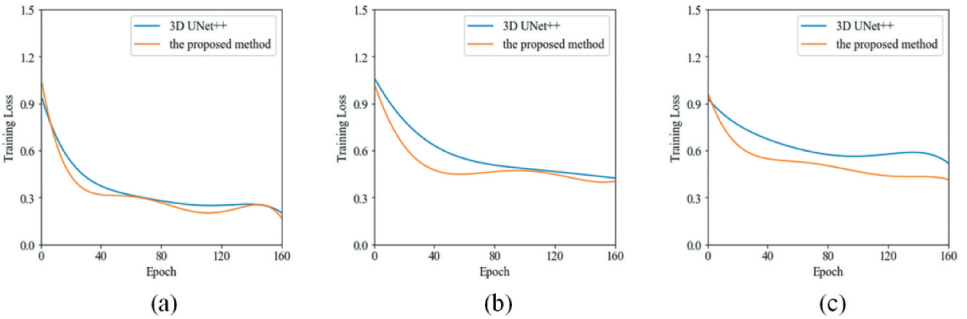

Figure 6. The training results of the 3D UNet++ and the proposed method over 160 epochs from (a) the T1-DUALin sequence, (b) the T1-DUALout sequence, and (c) the T2-SPIR sequence.

displays some segmentation results produced by using four methods. It can be found that there are more under-segmentation and over-segmentation in the results of the first three methods.

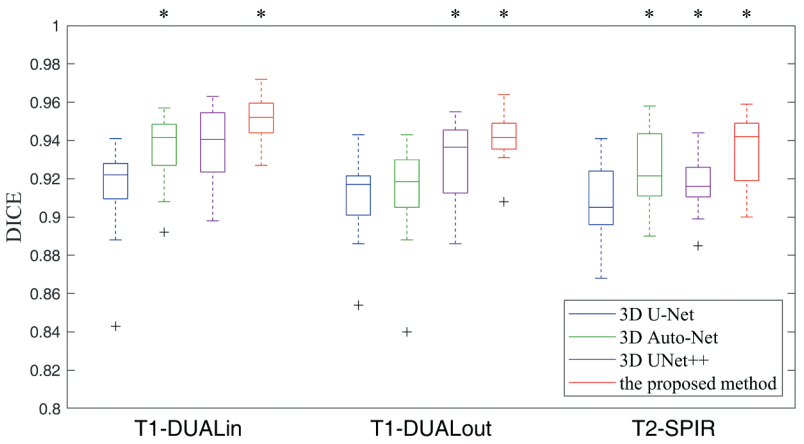

Figure 7. The boxplot of DICE results using four methods on multi-sequence MR images of the CHAOS19 dataset. A star indicates a statistical significant difference of the median DICE compared to the previous column.

Table 2. The mean and standard deviation of quantitative measures for liver segmentation on multi-sequence MR images of the CHAOS19 dataset.

	Methods	DICE	RAVD	ASSD(mm)	MSSD(mm)
T1-DUALin	3D U-Net	0.916±0.021	4.162±3.420	1.268±0.689	19.378±6.962
	3D Auto-Net	0.936±0.018	3.987±2.949	**0.711**±0.362	19.161±5.346
	3D UNet++	0.938±0.019	3.663±2.608	0.878±0.681	18.856±5.193
	the proposed method	**0.951**±0.012	**3.299**±1.603	0.834±0.625	**16.750**±4.328
T1-DUALout	3D U-Net	0.911±0.021	4.990±1.535	1.601±1.809	24.347±17.257
	3D Auto-Net	0.914±0.024	4.674±6.358	1.245±0.640	23.819±9.686
	3D UNet++	0.930±0.021	4.515±2.071	1.466±0.830	22.369±8.108
	the proposed method	**0.942**±0.012	**3.040**±1.621	**1.102**±1.101	**20.461**±5.989
T2-SPIR	3D U-Net	0.908±0.019	7.090±2.279	1.825±1.080	26.990±9.357
	3D Auto-Net	0.926±0.019	5.340±2.628	**1.285**±0.805	25.304±9.741
	3D UNet++	0.918±0.014	5.891±2.164	1.782±1.146	23.089±8.905
	the proposed method	**0.935**±0.018	**3.654**±2.651	1.397±0.639	**22.240**±10.163

Comparison with Transformer-Based Methods

For segmentation on all CT images, we compared the proposed method to the Swin-Unet method (Cao et al. 2022). In principle, Swin Transformer block that computes self-attention within 2D local windows (Liu et al. 2021) is taken as the basic unit of U-shaped architecture in this method. The Swin-Unet model was trained using the Adam optimizer for 100 epochs with a learning rate of 1*e*-4, a batch size of 8, patch size of 16, and pre-trained initial weights. In quantitative measures (see Table 3), the Swin-Unet method obtained the better MSSD than the proposed method.

For segmentation on all MR images, we compared the proposed method to the UNETR method (Hatamizadeh et al. 2022). In this method, the Transformer operating 3D input volumes is employed as the main encoder of network. The UNETR model was trained using the AdamW optimizer for

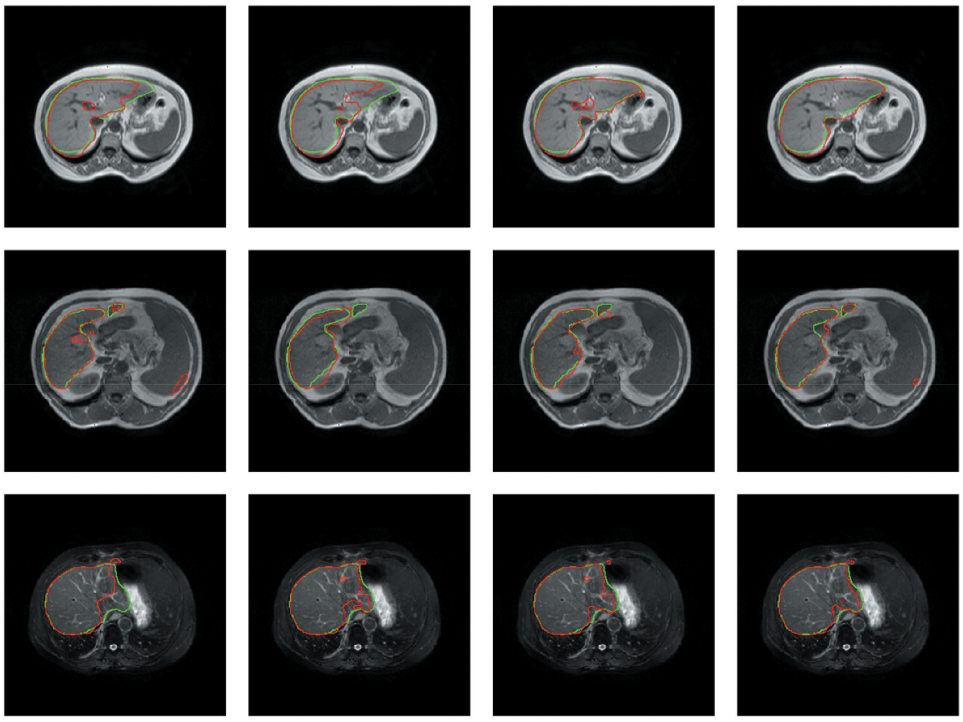

Figure 8. Typical liver segmentation results on MR images with the T1-DUALin (top row), the T1-DUALout (middle row), and the T2-SPIR (bottom row) sequences by using four methods. (a) 3D U-Net. (b) 3D Auto-Net. (c) 3D UNet++. (d) the proposed method. Green contours indicate the ground truth segmentation, and red contours indicate the automatic segmentation by the algorithm.

Table 3. The overall quantitative results of liver segmentation on CT and MR images as mean by using the Transformer-based methods and the proposed method.

	Methods	DICE	RAVD	ASSD(mm)	MSSD(mm)
CT	Swin-Unet	0.947	4.103	2.152	28.541
	the proposed method	0.958	3.081	1.353	30.445
MR	UNETR	0.950	3.684	1.276	18.722
	the proposed method	0.943	3.331	1.111	19.817

25,000 epochs with a learning rate of 1e-4, a batch size of 2, patch size of 16, and pre-trained initial weights. It can be seen from Table 3 that segmentation accuracy of the UNETR method is slightly higher than the proposed method in terms of the DICE and MSSD.

Ablation Study

We conducted an ablation study to evaluate the contribution of the shape context. For MR images with 3D application, node $X^{1,0}$ in Figure 2 was

Table 4. The overall quantitative results of liver segmentation on MR images as mean by using the proposed method with different scales.

Methods	DICE	RAVD	ASSD(mm)	MSSD(mm)
IterConv-SSC	0.933	4.035	1.326	20.563
IterConv-MSC	0.943	3.331	1.111	19.817

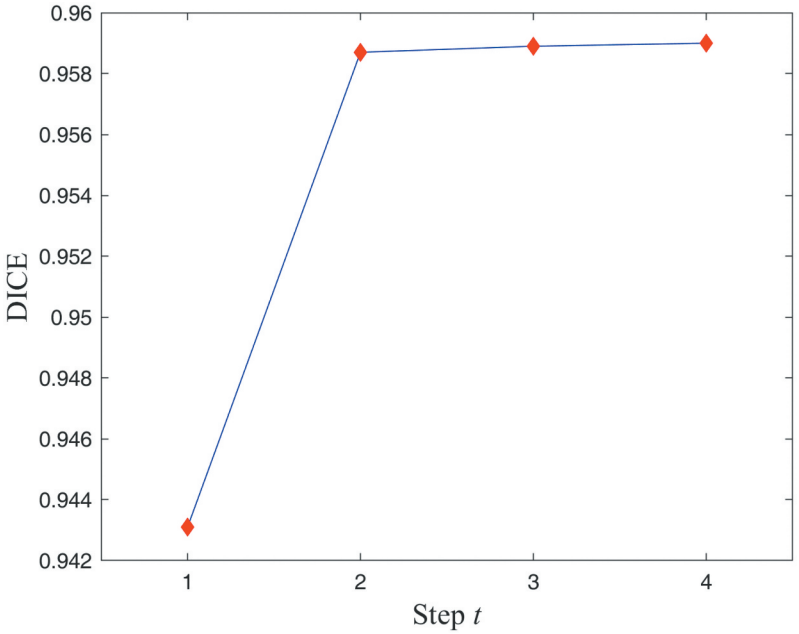

Figure 9. The DICE results of liver segmentation on all CT images as mean by using four steps of the proposed method.

replaced by the common convolutional block. It means that the posterior probabilities from the previous classifier are fused in a single-scale manner (IterConv-SSC). Except for changing of network structures, all parameters settings in this variation are the same as those in the proposed method in a multi-scale manner (IterConv-MSC). The overall results of quantitative measures for this variation are given in Table 4. They show that using IterConv-MSC brings an increase of DICE from 0.933 to 0.943, a decrease of RAVD from 4.035 to 3.331, a decrease of ASSD from 1.326 mm to 1.111 mm, and a decrease of MSSD from 20.563 mm to 19.817 mm. Therefore, the effectiveness of our method in a multi-scale manner is confirmed.

In addition, the influence of different steps ($t = 1, \ldots, K$) was examined. For CT images with 2D application, Figure 9 shows the mean DICE for liver segmentation at four steps of the proposed algorithm. It can be observed that the networks learned multi-scale context information through iterations for improvement in the Dice coefficient. Therefore, $K = 2$ is a good compromise between segmentation accuracy and computational burden.

Table 5. The model complexity of different networks.

Methods	Params	FLOPs
2D U-Net	1.95M	15.98G
2D Auto-Net	1.95M	16.02G
2D UNet++	2.29M	34.57G
the proposed method (2D)	2.32M	36.73G
3D U-Net	4.12M	167.58G
3D Auto-Net	4.12M	168.18G
3D UNet++	6.87M	664.63G
the proposed method (3D)	6.96M	697.08G

Model Complexity

The model complexity in terms of trainable parameters and floating-point operations (FLOPs) for the proposed method and various baseline methods is listed in Table 5. It shows that the proposed method still have an acceptable computational complexity, although containing the probabilistic map. In the inference stage of 2D application, it costs about 0.1 s to segment each CT slice. For 3D application, the inference time is about 1.3 s to segment each MR volume. Since the model can be trained offline, our method would be practicable and efficient in routine clinical workflow.

Conclusion

In this work, we proposed an iterative convolutional encoder-decoder network, which integrates multi-scale context information for liver segmentation. We evaluated this model on abdominal CT and MR images of three public datasets. The experimental results show that the proposed model is able to produce more accurate liver segmentation than other models. In future work, we will embed attention mechanisms into this model for further improvement (Zhang et al. 2022). Moreover, the initial probabilistic map from multi-atlas registration (Zhang et al. 2021) and segmentation post-processing (Chen et al. 2022) would be another future direction.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 61002046, and in part by the Natural Science Foundation of Hubei Province of China under Grant 2016CFB489.

References

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12):2481–95. doi:10.1109/TPAMI.2016.2644615.

Bilic, P., Christ, P.F. *et al* . 2022 doi:. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*.

Cao, H. , Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. 2022. Swin-Unet: Unet-like pure transformer for medical image segmentation. Proceedings of the European Conference on Computer Vision Workshops Tel Aviv, Israel, 1–14.

Cerrolaza, J. J., M. Reyes, R. M. Summers, M. Á. González-Ballester, and M. G. Linguraru. 2015. Automatic multi-resolution shape modeling of multi-organ structures. *Medical Image Analysis* 25 (1):11–21. doi:10.1016/j.media.2015.04.003.

Chartrand, G., T. Cresson, R. Chav, A. Gotra, A. Tang, and J. A. De Guise. 2017. Liver segmentation on CT and MR using laplacian mesh optimization. *IEEE Transactions on Biomedical Engineering* 64 (9):2110–21. doi:10.1109/TBME.2016.2631139.

Chen, Y., W. Chen, X. Yin, X. Ye, X. Bao, L. Luo, Q. Feng, Y. Li, and X. Yu. 2011. Improving low-dose abdominal CT images by weighted intensity averaging over large-scale neighborhoods. *European Journal of Radiology* 80 (2):E42–49. doi:10.1016/j.ejrad.2010.07.003.

Chen, S., Gamechi, Z.S., Dubost, F., Tulder, G., Bruijne, M. 2022. An end-to-end approach to segmentation in medical images with CNN and posterior-CRF. *Medical Image Analysis* 76: 102311.

Chen, J., Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou. 2021. TransUnet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, H., X. Pan, X. Lu, and Q. Xie. 2020. A modified graph cuts image segmentation algorithm with adaptive shape constraints and its application to computed tomography images. *Biomedical signal processing and control* 62:102092. doi:10.1016/j.bspc.2020.102092.

Chen, J., Yang, L., Zhang, Y., Alber, Mark, Chen, D.Z. 2016 .Combining fully convolutional and recurrent neural networks for 3D biomedical image segmentation. Conference on Neural Information Processing Systems Centre Convencions Internacional Barcelona, Barcelona SPAIN, vol. 29, 1–9.

Cicek, O., Abdulkadir, A., Lienkamp, S., Brox, T., Ronneberger, O. 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. *arXiv preprint arXiv:1606.06650v1*.

Dumoulin, V., and F. Visin. 2018. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285v2*.

Gibson, E., F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. Davidson, S. P. Pereira, M. J. Clarkson, and D. C. Barratt. 2018. Automatic multi-organ segmentation on abdominal CT with dense v-networks. *IEEE Transactions on Medical Imaging* 37 (8):1822–34. doi:10.1109/TMI.2018.2806309.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D. 2022. UNETR: Transformers for 3d medical image segmentation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Waikoloa, HI, USA.

Heimann, T., B. van Ginneken, M. A. Styner, Y. Arzhaeva, V. Aurich, C. Bauer, A. Beck, C. Becker, R. Beichel, G. Bekes, et al. 2009. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Transactions on Medical Imaging* 28 (8):1251–65. doi:10.1109/TMI.2009.2013851.

Jin, C., F. Shi, D. Xiang, X. Jiang, B. Zhang, X. Wang, W. Zhu, E. Gao, and X. Chen. 2016. 3D fast automatic segmentation of kidney based on modified AAM and random forest. *IEEE Transactions on Medical Imaging* 35 (6):1395–407. doi:10.1109/TMI.2015.2512606.

Kavur, A. E., Gezer, N.S., Barış, M. *et al* 2020. CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. *arXiv preprint arXiv:2001.06535*.

Ker, J., L. Wang, J. Rao, and T. Lim. 2018. Deep learning applications in medical image analysis. *IEEE Access* 6:9375–89. doi:10.1109/ACCESS.2017.2788044.

Kingma, D. P., and J. Ba. 2014. ADAM: A method for stochastic optimization. [Online]. https://arxiv.org/abs/1412.6980.

Li, G., X. Chen, F. Shi, W. Zhu, J. Tian, and D. Xiang. 2015. Automatic liver segmentation based on shape constraints and deformable graph cut in CT images. *IEEE Transactions on Image Processing* 24 (12):5315–29. doi:10.1109/TIP.2015.2481326.

Litjens, G., T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* 42:60–88. doi:10.1016/j.media.2017.07.005.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030v2*.

Long, J., Shelhamer, E., Darrell, T. 2015. Fully convolutional networks for semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition Boston, MA, USA, 3431–40.

Moghbel, M., S. Mashohor, R. Mahmud, and M. I. B. Saripan. 2018. Review of liver segmentation and computer assisted detection/diagnosis methods in computed tomography. *Artificial Intelligence Review* 50 (4):497–537. doi:10.1007/s10462-017-9550-x.

Nair, V., and G. E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. Proceedings of ICML Haifa, Israel, 807–14.

Oktay, O., Schlemper, J., Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N., Kainz, B., Glocker, B., Rueckert, D. 2018. Attention U-Net: Learning where to look for the pancreas. Conference on Medical Imaging with Deep Learning Amsterdam, 1–10.

Paszke, A., Gross, S., Chintala, S. *et al* 2017. Automatic differentiation in pytorch. *NIPS-W*.

Ronneberger, O., Fischer, P., Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention Munich, Germany, 234–41.

Salehi, S. S. M., D. Erdogmus, and A. Gholipour. 2017. Auto-context convolutional neural network for brain extraction in magnetic resonance imaging. *IEEE Transactions on Medical Imaging* 36 (11):2319–30. doi:10.1109/TMI.2017.2721362.

Shen, D., G. Wu, and H.-I. Suk. 2017. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering* 19 (1):221–48. doi:10.1146/annurev-bioeng-071516-044442.

Tang, A., J. Chen, T.-A. Le, C. Changchien, G. Hamilton, M. S. Middleton, R. Loomba, and C. B. Sirlin. 2015. Cross-sectional and longitudinal evaluation of liver volume and total liver fat burden in adults with nonalcoholic steatohepatitis. *Abdominal imaging* 40 (1):26–37. doi:10.1007/s00261-014-0175-0.

Wang, J., Y. Cheng, C. Guo, Y. Wang, and S. Tamura. 2016. Shape–intensity prior level set combining probabilistic atlas and probability map constrains for automatic liver segmentation from abdominal CT images. *International Journal of Computer Assisted Radiology and Surgery* 11 (5):817–26. doi:10.1007/s11548-015-1332-9.

Yang, X., L. Yu, S. Li, H. Wen, D. Luo, C. Bian, J. Qin, D. Ni, and P.-A. Heng. 2019. Towards automated semantic segmentation in prenatal volumetric ultrasound. *IEEE Transactions on Medical Imaging* 38 (1):180–93. doi:10.1109/TMI.2018.2858779.

Zhang, Y., He, Z., Zhong, C., Zhang, Y., Shi, Z. 2017. Fully convolutional neural network with post-processing methods for automatic liver segmentation from CT. *Chinese Automation Congress*, 3864–69.

Zhang, Y., Jiang, X., Zhong, C., Zhang, Y., Shi, Z., Li, Z. 2018. SequentialSegNet: Combination with sequential feature for multi-organ segmentation. International Conference on Pattern Recognition Beijing,China, 3864–69.

Zhang, Y., J. Wu, Y. Liu, Y. Chen, W. Chen, E. Wu, C. Li, and X. Tang. 2021. A deep learning framework for pancreas segmentation with multi-atlas registration and 3D level-set. *Medical Image Analysis* 68:101884. doi:10.1016/j.media.2020.101884.

Zhang, Y., Yang, J., Liu, Y., Tian, J., Wang, S., Zhong, C., Shi, Z., Yang, Z., He, Z. 2022. Decoupled pyramid correlation network for liver tumor segmentation from CT images. *Medical Physics*: 1–15. doi:10.1016/j.ejmp.2022.04.018.

Zhou, Z., Siddiquee, M., Tajbakhsh, N., Liang, J. 2020. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging* 37 (8):1822–34.