Computers in Biology and Medicine 101 (2018) 153–162

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Robust liver vessel extraction using 3D U-Net with variant dice loss function

Qing Huang[a], Jinfeng Sun[a], Hui Ding[a], Xiaodong Wang[b], Guangzhi Wang[a,*]

[a] Department of Biomedical Engineering, School of Medicine, Tsinghua University, Room C249, Beijing, 100084, China
[b] Department of Interventional Radiology, Peking University Cancer Hospital & Institute, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Beijing, 100142, China

ARTICLE INFO

Keywords:
Liver vessel extraction
3D U-Net
Variant dice loss function
Annotation quality
Refined manual expert annotations

ABSTRACT

Purpose: Liver vessel extraction from CT images is essential in liver surgical planning. Liver vessel segmentation is difficult due to the complex vessel structures, and even expert manual annotations contain unlabeled vessels. This paper presents an automatic liver vessel extraction method using deep convolutional network and studies the impact of incomplete data annotation on segmentation accuracy evaluation.

Methods: We select the 3D U-Net and use data augmentation for accurate liver vessel extraction with few training samples and incomplete labeling. To deal with high imbalance between foreground (liver vessel) and background (liver) classes but also increase segmentation accuracy, a loss function based on a variant of the dice coefficient is proposed to increase the penalties for misclassified voxels. We include unlabeled liver vessels extracted by our method in the expert manual annotations, with a specialist's visual inspection for refinement, and compare the evaluations before and after the procedure.

Results: Experiments were performed on the public datasets Sliver07 and 3Dircadb as well as local clinical datasets. The average dice and sensitivity for the 3Dircadb dataset were 67.5% and 74.3%, respectively, prior to annotation refinement, as compared with 75.3% and 76.7% after refinement.

Conclusions: The proposed method is automatic, accurate and robust for liver vessel extraction with high noise and varied vessel structures. It can be used for liver surgery planning and rough annotation of new datasets. The evaluation difference based on some benchmarks, and their refined results, showed that the quality of annotation should be further considered for supervised learning methods.

1. Introduction

Computer-assisted liver interventional surgery (e.g. ablation and embolization) is an effective treatment for unresected liver tumors. Needle-shaped instruments need to be safely inserted into a target area without penetrating liver vessels in liver biopsy or ablation [1]. Liver vessels that feed the tumors should be accurately located in liver embolization surgery for planning. The relative position between liver vessels and liver tumor and information like vessel diameter may determine ablation results and affect local tumor recurrence rates [2]. CT liver vessel extraction is essential for 3D visualization, path planning and guidance in liver interventional surgery. Manual delineation of liver vessels on each slice is both time consuming and error-prone, and results are inconsistent between different experts. Since annotation is very tedious, some vessels may be unlabeled. Therefore, an automatic, accurate and robust liver vessel extraction algorithm is clearly needed in clinical settings.

Automatic liver vessel extraction is still challenging due to the complicated and diverse structure of liver vessels (different shape, size, length and positions). The low contrast with surrounding tissues, high noise and irregular vessel shapes caused by nearby tumors all increase the difficulty of liver vessel segmentation [3].

Current liver vessel segmentation methods can be roughly classified into image filtering and enhancement algorithms, deformable model-based algorithms, tracking-based algorithms and machine learning-based algorithms [4]. In image filtering and enhancement methods (e.g. Hessian-based filters, Gabor filters and diffusion filters), the vessels were usually enhanced by utilizing image gradients or multiscale high order deviations, then extracted by region-growing, graph cuts and so on [5–8]. A review of image filtering methods can be found in Refs. [9,10]. While these methods were effective, they may not do well on images with high noise or non-uniform intensities and required complex parameter adjustment [10,11]. Deformable model-based algorithms (like level set) were usually sensitive to initial seeds or contour

* Corresponding author.
E-mail addresses: q-huang12@mails.tsinghua.edu.cn (Q. Huang), sjf16@mails.tsinghua.edu.cn (J. Sun), dinghui@tsinghua.edu.cn (H. Ding), tigat@126.com (X. Wang), wgz-dea@tsinghua.edu.cn (G. Wang).

https://doi.org/10.1016/j.compbiomed.2018.08.018
Received 22 March 2018; Received in revised form 17 August 2018; Accepted 17 August 2018

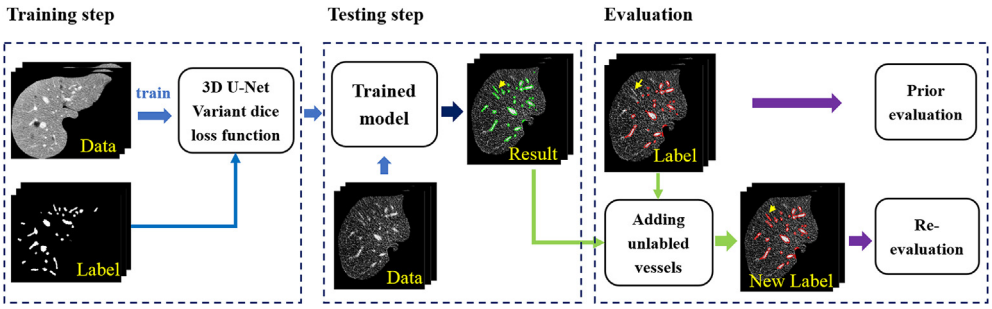

Fig. 1. The flowchart of the proposed method.

selection and parameters, they may leak into adjacent tissues in low contrast images [12,13]. Tracking-based algorithms usually include model-based algorithms that track the vessels based on the predefined vessel models and tracking methods based on the minimum cost path [14–16]. While these algorithms are prone to errors in vessels with irregular shape near liver tumors and user interaction was usually needed [14,16,17].

In machine classification-based algorithms, the vessel was first extracted by k-means clustering, and refined iteratively using linear contrast stretching and morphological operations for vessel reconstruction [18]. Vessel features like intensity value, vessel saliency, direction and connectivity were first extracted and used for liver vasculature grouping, then multiple feature point voting was implemented for further grouping [3]. Another algorithm used four different typical vessel filters for liver vessel feature extraction, and applied extreme learning machine for vessel classification [19]. In the previous methods, authors need to carefully design the characteristics of liver vessels and balance different parameters.

Recently, deep convolutional networks have achieved remarkable results in medical image segmentation. These networks can automatically learn complex image features and combine them into hierarchical representations for prediction and classification. However, there are still difficulties of deep learning methods for liver vessel extraction. The accuracy of deep learning methods relies upon the number and variety of training samples. For liver vessels, the structures are complex and diverse. There are only limited training samples for 3D medical datasets with annotations. Moreover, the foreground and background class voxels are unbalanced. Furthermore, some liver vessels were missed in the manual annotations (even though the annotation was done by experienced experts with few biases) and the benchmark for training and evaluation was incomplete. Actually, this is a great concern for the quality of annotation in medical datasets, since the clinical experience of the annotators might be quite different, and annotation results differ. Using annotations without considering reliability as the benchmark for supervised learning methods may cause segmentation bias.

Kitrungrotsakul et al. [20] applied three deep convolutional networks with shared kernels to extract features from different views of CT images. Weighted parameters should be selected for the log-likelihood loss function and the network failed in datasets with different intensity distributions than the trained models [20]. Furthermore, their evaluation was based on the reference datasets with incomplete annotations.

3D U-Net is a full and dense convolutional network that can achieve end-to-end image segmentation with few training samples and incomplete annotations [21]. Unbalanced class frequency often causes the learning process to become trapped in a local minima of the loss function, and leads to predictive bias. Weighted parameters of different classes need to be pre-computed to compensate for different class frequency in U-Net [22]. A novel objective loss function based on the dice coefficient was proposed to deal with class imbalance without establishing the right balance between foreground and background classes [23]. Then, the Tversky loss function, which is a variant of the dice coefficient made by adjusting the parameters of over- or under-segmented foreground pixel numbers, was proposed and achieved more accurate results than the method with dice loss function in lesion segmentation [24]. However, the algorithm still needs to balance segmentation accuracy (dice value) and sensitivity with different parameters.

In this paper, we choose the 3D U-Net for automatic liver vessel extraction with few training samples and incomplete labeling. To adapt to the problems with high class imbalance and improve segmentation accuracy, we adjust the penalties for the number of misclassified voxels and propose a new similarity metric (a variant of dice coefficient) for the loss function. The annotations are refined by including un-annotated liver vessels that were extracted by our deep-learning method. Then the difference between the evaluation with original annotations and with refined annotations is compared to show the impact of incomplete or inaccurate annotations on segmentation accuracy evaluation. The proposed algorithm is proved to be automatic, accurate and robust for liver vessel extraction even for images with high noise, low contrast and varied vessel structures, and superior to the method using the dice loss function. The algorithm can efficiently build the relative relationship of liver tumor and vessels in 3D and be used for liver surgery planning. The phenomenon of evaluation difference caused by different benchmarks should also be noted by researchers, so as to take care concerning medical annotation quality, in order to obtain results that are closer to actual values.

2. Methods

The proposed method starts with data preprocessing, then applies 3D U-Net for liver vessel enhancement and classification, and post-processing for refinement. Finally, the existing reference datasets are refined by including unlabeled vessels that were detected by our method, and the method is reevaluated with the new reference for comparison. Fig. 1 shows the flowchart of the algorithm.

2.1. Training and testing datasets

3Dircadb datasets (https://www.ircad.fr/research/3d-ircadb-01/) are currently available public datasets with liver and liver vessel contours for training and evaluation of liver vessel extraction algorithms [20,25,26]. The datasets include 20 contrast-enhanced CT volumes with various image resolutions, vessel structures, intensity distributions and contrast between liver and liver vessels. In the experiment, we selected 10 cases from 3Dircadb datasets with annotated data as the training data, and another 10 cases as the testing data. We also tested our algorithm on 20 Sliver07 CT datasets (http://www.sliver07.org/) and 10 local CT datasets from Peking University Cancer Hospital.

Considering the accuracy, invariance and robustness of the network, most of the pre-selected training datasets should have clear, abundant and varied liver vessel structures, different intensity ranges and contrast between liver and liver vessels. The liver vessel appearances should be similar in both the training datasets and testing datasets (both include

easy and difficult cases). 10 experiments were performed, and the training and testing datasets were selected based on the above metrics by hand in each experiment. The training and testing datasets within one experiment that had both better performances and small evaluation difference were selected as the datasets for our experiment.

Rotation and mirroring operations were performed on the training samples for data augmentation to use the available annotated data efficiently. In the training data, pixel spacing varied from 0.57 to 0.87 mm, slice thickness varied from 1 to 4 mm and slice number varied from 74 to 260. In the test data, pixel spacing varied from 0.56 to 0.81 mm, slice thickness varied from 1.25 to 2 mm and slice number varied from 113 to 225 in the 3Dircadb datasets. Pixel spacing varied from 0.58 to 0.81 mm, slice thickness varied from 0.7 to 5 mm and slice number varied from 64 to 394 in the Sliver07 datasets. Pixel spacing varied from 0.68 to 0.98 mm, slice thickness varied from 0.63 to 1 mm and slice number varied from 353 to 713 in the local datasets.

2.2. Preprocessing

Preprocessing consists of 3 steps: (1) CT values are limited to [0400] HU to focus on the intensity range of the liver. (2) CT images and annotated images are cropped to the liver area based on the pre-segmented liver mask, and adjusted to the size of $288 \times 288 \times 96$. For local datasets, the liver mask is obtained by our previous liver segmentation method [27]. (3) Images are normalized to zero mean and unit variance. Since most liver vessels are quite small, we trained images with their original resolution if it was in a reasonable range to prevent artifact errors caused by resampling. For data with slice thickness smaller than 0.8 mm or larger than 2.5 mm, we used a coordinate transform and cubic spline interpolation to transfer the data into a slice thickness of 1.6 mm.

2.3. 3D U-Net architecture

3D U-Net is a dense network that can be used for volumetric segmentation with sparsely annotated training data [21]. Fig. 2 illustrates our proposed network architecture based on 3D U-Net for liver vessel extraction. It contains an encoding path (left side) for feature extraction and a decoding path (right side) for full-resolution segmentation. The encoding path follows the convolutional network and has five resolution layers. Each layer has two padded convolutions (size $7 \times 7 \times 7$ on first layer and $3 \times 3 \times 3$ on other layers), followed by a batch normalization (BN) step to prevent bottlenecks of network architecture [28], a rectified linear unit (ReLU) as an activation function, and a 2x2x2 max pooling operation with stride 2 for down-sampling and over-fitting prevention. For accurate and sufficient deep feature extraction, the network doubles the number of feature channels (denoted beside each feature map as shown in Fig. 2 starting from 16) along each successive layer in the encoding path. In the decoding path, each layer has a $2 \times 2 \times 2$ up-convolution with stride 2 for up-sampling, followed by two padded convolutions (size $7 \times 7 \times 7$ on last layer and $3 \times 3 \times 3$ on other layers), a BN step and a ReLU. The number of feature channels is halved along each successive layer in the decoding path. The last convolution at the last layer without BN is used to map the final 16-component feature vector to the desired number of feature channels. At each layer with the same resolution, a concatenated path is built by passing the corresponding cropped feature map from the encoding path to decoding path to prevent accuracy decreasing caused by lost border pixels. A softmax function is finally used for classification.

The input data of the network is a 3D patch that is randomly picked from all training samples. Larger patches provide more abundant contexts and broader features, but also need more computation resource. Considering the desired segmentation accuracy and the performance of our computer, the patch size was set to $96 \times 96 \times 96$ and batch size was set to 3 in the experiment. The typical Adam optimizer [29] is selected for network training. The initial learning rate was 0.001 and the maximum number of iterations was 21500.

2.4. Loss function

Liver vessels occupy only a small portion of the liver, and unbalanced foreground (liver vessel) and background (liver) classes often cause predictive deviation and bias the classification to the background with more voxels. Considering imbalance problems and class weight parameter adjustment problems, the similarity metric of the dice coefficient which evaluates the segmentation accuracy was proposed for the loss function [23]. The dice coefficient is calculated as follows:

$$Dice(P, G) = \frac{|P \cap G|}{|P \cap G| + 0.5(|P - G| + |G - P|)} \quad (1)$$

Where P is the predicted labels, G is the labels of the ground truth (i.e., the annotated data). $|P \cap G|$ represents the number of correctly classified foreground voxels, i.e., true positive (T_p). $|P - G|$ represents the number of misclassified background voxels to foreground voxels, i.e.,

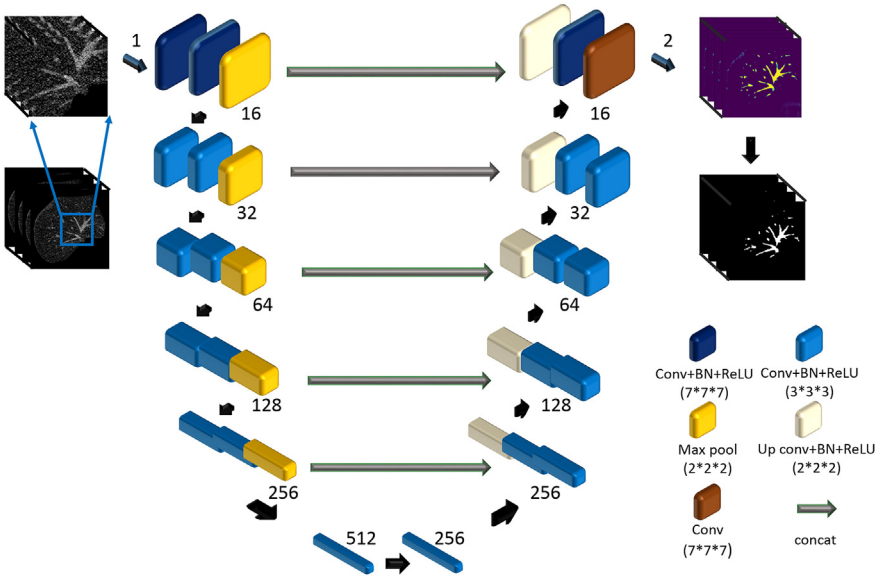

Fig. 2. The proposed network architecture for liver vessel extraction. The number of feature channels is defined beside each feature map.

false positive or over-segmented voxels. $|G - P|$ represents the number of misclassified foreground voxels to background voxels, i.e., false negative or under-segmented voxels. And $|P - G| + |G - P|$ represents the number of all misclassified voxels.

To adapt to unbalanced class frequency but also improve segmentation accuracy, we adjust the penalty weights of the misclassified voxels to get a higher T_p value and lower the number of misclassified voxels. A new similarity metric $M(P, G, \beta)$, which is a variant of the dice coefficient, is proposed for the lose function:

$$M(P, G, \beta) = \frac{|P \cap G|}{|P \cap G| + 0.5\beta(|P - G| + |G - P|)} \quad (2)$$

Where parameter β determines the weight of the number of correctly classified foreground voxels and the number of misclassified voxels.

The class number is 2 in our situation, and we take the foreground and background as the first class and second class respectively. Then $M(P, G, \beta)$ in (2) is calculated as follows:

$$M(\beta) = \frac{\sum_{i=1}^{N} p_{0i} g_{0i}}{\sum_{i=1}^{N} p_{0i} g_{0i} + 0.5\beta(\sum_{i=1}^{N} p_{0i} g_{1i} + \sum_{i=1}^{N} p_{1i} g_{0i})} \quad (3)$$

Where p_{0i} and p_{1i} are the probabilities that voxel i belongs to the foreground (liver vessel) and the background (liver) respectively in the softmax layer output result. g_{0i} and g_{1i} are the labels of voxel i in the annotated data for liver vessel or liver respectively with value 0 or 1.

The gradient of the similarity in (3) with respect to p_{0i} and p_{1i} are calculated as follows:

$$\frac{\partial M}{\partial p_{0i}} = \frac{g_{0j}(\sum_{i=1}^{N} p_{0i} g_{0i} + 0.5\beta(\sum_{i=1}^{N} p_{0i} g_{1i} + \sum_{i=1}^{N} p_{1i} g_{0i})) - \sum_{i=1}^{N} p_{0i} g_{0i}(g_{0j} + 0.5\beta g_{1j})}{(\sum_{i=1}^{N} p_{0i} g_{0i} + 0.5\beta)(\sum_{i=1}^{N} p_{0i} g_{1i} + \sum_{i=1}^{N} p_{1i} g_{0i})^2} \quad (4)$$

$$\frac{\partial M}{\partial p_{1i}} = \frac{-\sum_{i=1}^{N} p_{0i} g_{0i}(g_{0j} + 0.5\beta g_{0j})}{(\sum_{i=1}^{N} p_{0i} g_{0i} + 0.5\beta)(\sum_{i=1}^{N} p_{0i} g_{1i} + \sum_{i=1}^{N} p_{1i} g_{0i})^2} \quad (5)$$

With these equations, the weights of the liver and liver vessel classes do not need to be pre-computed and balanced. The proposed algorithm can adjust the penalty for misclassified foreground voxels and improve the classification accuracy by selecting a proper β experimentally.

2.5. Post-processing

In post-processing, region connectivity is first performed on the vessels extracted by 3D U-Net. To remove some noises caused by classification, regions with small volumes (less than 120 mm^3) are removed (see Fig. 3 (d) (e)). To remove some misclassifications, regions unconnected and far away from large vessels with volumes less than 320 mm^3 are eliminated. To prevent the impact of bright lipiodol, regions with pixel intensities higher than 350 HU are removed too.

2.6. Refining annotated datasets for comparison

Complicated liver vessel structures, high noise, low contrast and fuzzy boundaries all add difficulties to manual segmentation. In our study, we found some vessels are unlabeled in the annotated datasets even though the vessels were annotated by an expert (as pointed out by the arrow in Figs. 10 and 11). The results of supervised deep-learning method showed some interesting phenomenon that they can detect vessels that unlabeled in the learned annotations. Based on the result, we extract the unlabeled vessels and add them to the annotated dataset for refinement, re-evaluation and comparison. We define Seg_m as our segmentation result and Seg_r as the expert annotated dataset. The refined procedure is below:

(1) Extracting the overlapping volumes of Seg_m and Seg_r. To prevent the influence of bordering pixels, we calculate the overlap rate of an object in each slice of Seg_m to each object in the corresponding slice of Seg_r, and keep objects of Seg_m with an overlap rate larger than 0.35.
(2) Extracting the relative over-segmented areas of Seg_m to Seg_r, i.e., Seg_m without (1).
(3) Manually selecting vessel voxels from (2) by an interventional surgeon with 12 years of experience and adding them to Seg_r for refinement.

3. Experiments and results

3.1. Evaluation metrics and platform

Four segmentation evaluation metrics were selected to evaluate the proposed method. These metrics include the dice coefficient for overall segmentation accuracy evaluation, voxel segmentation accuracy, sensitivity and specificity [30]. The accuracy was estimated by the ratio of the total number of correctly classified foreground and background voxels to the number of voxels in the liver volume. Sensitivity indicated the number of properly classified foreground voxels with respect to the number of foreground voxels, and specificity indicated the number of properly classified background voxels with respect to the number of background voxels. Each of the four metrics highlighted one different aspect of segmentation quality and was used for a general evaluation.

Our proposed method was implemented using python 3.6 and Tensorflow. The program was implemented on a computer with an Intel Xeon(R) E5620 CPU (@ 2.4 GHz, 23.5 GB RAM) and an NVIDIA Quadro M4000 GPU (8 GB memory). The training time of our model was about 48 h, and the testing time for 20 3Dircadb datasets was between 2.04 and 8.37 min with average testing time 3.8 min.

3.2. Comparisons of loss function

The loss function in our model was based on a variant of the dice coefficient, and parameter β in (2) was integrated into the proposed similarity metric. Fig. 4 shows the average dice, accuracy and sensitivity of the proposed network without post-processing on 20 3Dircadb datasets with β ranging from 1 to 8. A change in β could affect the weights of the number of correctly classified foreground voxels and the number of misclassified voxels. Generally, a smaller β corresponds to a larger degree of optimization for correctly classified foreground voxels, but a weaker penalty for misclassified voxels. A larger β has the

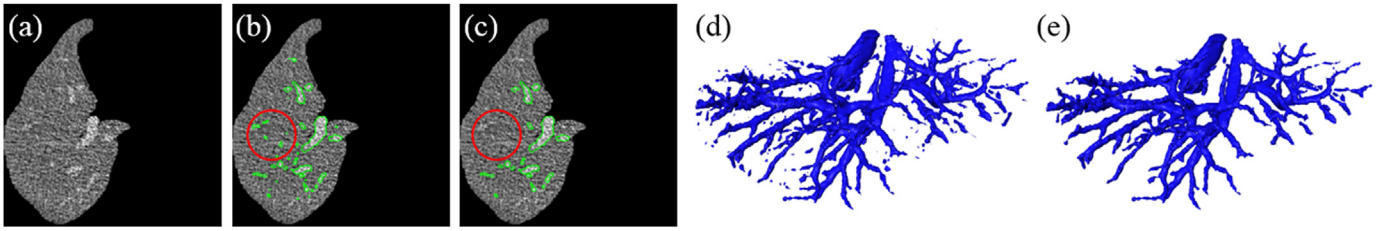

Fig. 3. The process of post-processing. (a) CT image; (b) results before post-processing; (c) removing some non-vessel areas after post-processing; (d) (e) 3D result before and after post-processing.

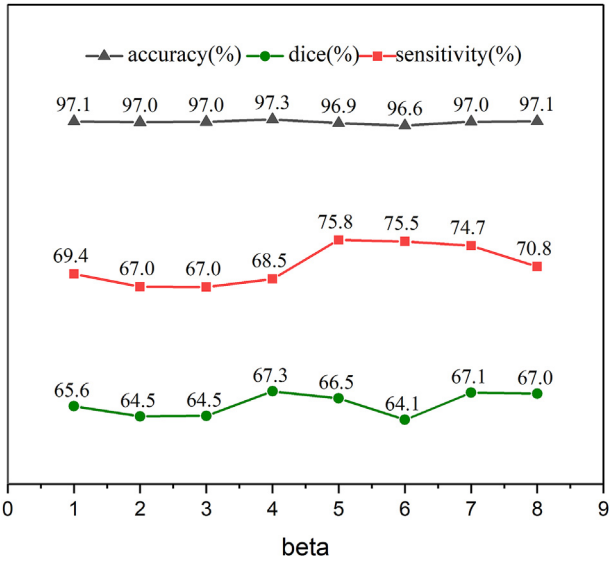

Fig. 4. Evaluation of the proposed network on 20 3Dircadb datasets with different β value.

opposite properties, and it can slow down the gradient of the loss function at the beginning and may cause the loss function to become trapped in a local minimum. Considering the segmentation accuracy and sensitivity, β was set to 7 in the experiment. The average dice value and sensitivity of our method were 67.1% and 74.7% respectively, compared to 65.6% and 69.4 for the network using the dice loss function ($\beta = 1$).

Fig. 5 shows the performance of the proposed loss function with different β on liver vessel extraction. To better explain the effect of parameter β in learning and classification, the results were not post-processed too. As pointed out by the yellow arrows, some labelled or unlabeled liver vessels were not detected by the network with the typical dice loss function ($\beta = 1$). As β grows, the penalty for misclassified liver vessels grows, and more liver vessels can be detected. Some β may increase the true positive and false positive at the same time. If β is too large, the optimization ability of the network decreases, and some under- or over-segmentation happened (see the arrows in Fig. 5). Good performances achieved when β was in the range of 5–7, and a proper β ($\beta = 7$) was chosen to balance different properties of the trained model for all datasets.

Fig. 6 shows the comparative results of our network using dice and our proposed loss function on different CT datasets. It was hard to describe and extract the features of some vessels with weak and fuzzy boundaries, or different intensity distribution from nearby vessels (see Fig. 6 (b) (f)). The network using the dice loss function had failed to learn and extract these features. With our modification, the learning ability for these easily misclassified vessels was enhanced, and the extracted vessel contours were closer to the real vessel boundaries. The vessels in the images with high noise, low contrast and varied structures were more accurately extracted by our method than the dice loss function method.

3.3. Sensitivity to noise evaluation

Segmentation of liver vessels from noisy images is a very challenging task since most of the vessels are thin and elongated, and high noise can severely blur boundaries and decrease contrast. Generally, the smaller the slice thickness of a CT volume, the higher the image noise. We chose one typical CT liver dataset as the synthetic dataset and added Gaussian white noise with different levels of noise variances to the dataset. Experiments were performed on these synthetic datasets to evaluate the sensitivity of our method to noise. Fig. 7 shows the synthetic data with different additive noise variances (from left to right: 0, 20, 50 and 80 HU) and the corresponding performance of our algorithm. Fig. 8 shows the evaluation result of the proposed method on the synthetic data with different added noise variances (from 0 to 90 HU).

As seen in Fig. 7, with the increase of added noise variance, some small and thin vessels were greatly spoiled by noise. Small or weak vessels can be detected by our algorithm with small additive noise, but were hard to extract with large additive noise. Medium or large size liver vessels can be effectively extracted with different additive noise variances. As shown in Fig. 8, the dice value ranged from 62.5 to 65.2%, sensitivity ranged from 87.8% to 90.7% and accuracy ranged from 97.4% to 97.7% with the noise variance amplitude smaller than 60 HU. The change of the dice, sensitivity and accuracy were all smaller than 3%. Thus, the algorithm is robust for assessing data with varied noise levels in an accepted range.

3.4. Results on slices with separated liver

In some slices, the liver is disjointed. For most training datasets, slices with partitioned livers only occupy a small proportion of the dataset. Fig. 9 shows the performance of our method on slices with separated liver partitions from a local dataset. The extracted liver vessel contours were close to the real contours, and vessels with different sizes or shapes on different liver partitions can be detected. The method proved to be accurate and robust for datasets with detached liver parts after the visual inspection by the specialist with 12 years of experience. The results of local datasets were found to be satisfying and acceptable by two specialists with 12 and 15 years of experience respectively.

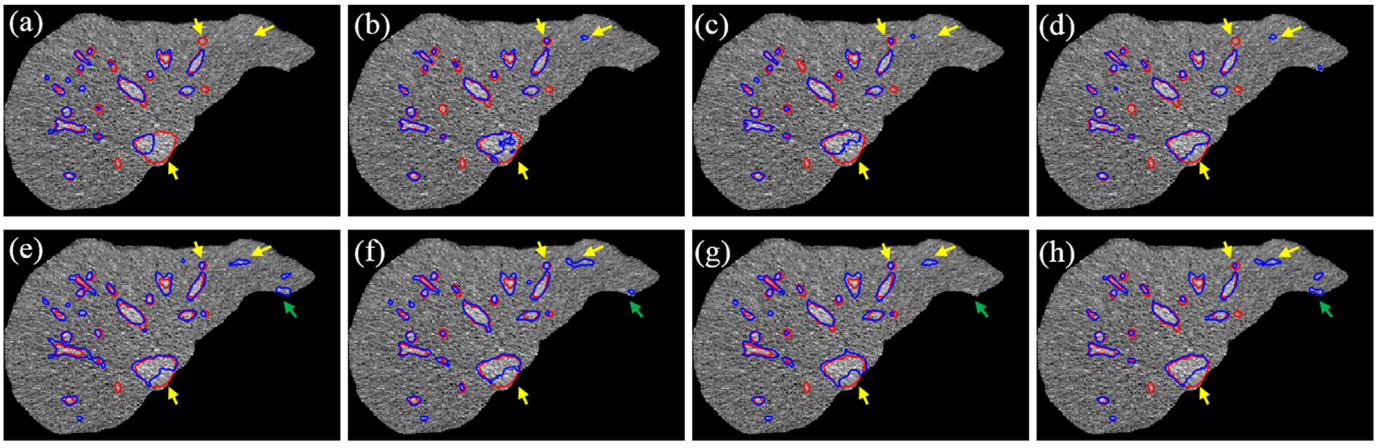

Fig. 5. An example of performance of the proposed loss function with different β values ((a)–(h): β from 1 to 8).

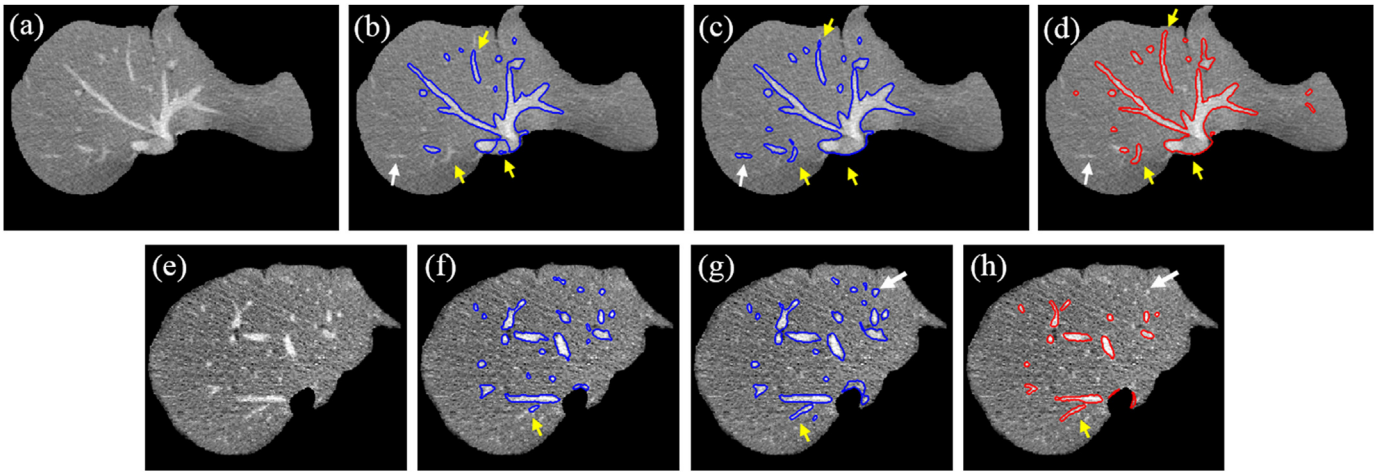

Fig. 6. Examples of performance obtained by the network with dice loss function ((b) (f)) and our loss function ((c) (g)). (a) (e): CT images; (d) (h): segmentation result by expert. Yellow arrows point to different extracted vessels and white arrows point to unlabeled vessels in the annotated data that were extracted by our method. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

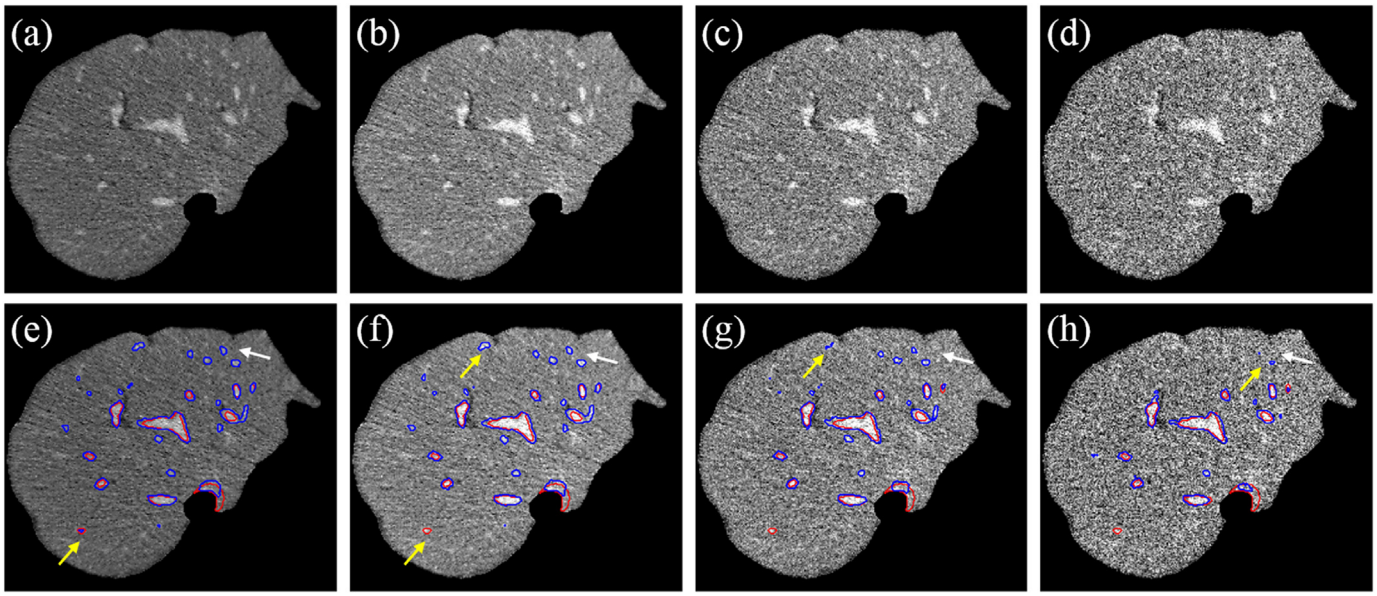

Fig. 7. Synthetic data with different noise variances (variances amplitude from left to right was 0, 20, 50 and 80 HU) and the corresponding performance by our algorithm (blue line) and expert (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.5. Evaluation comparison with unrefined and refined benchmarks

The quality of medical dataset annotation can affect the training and evaluation of the segmentation algorithms. We compared the segmentation and evaluation results with the original annotation and with refined annotation as benchmarks. We calculated the number of the detected unlabeled vessel voxels by our algorithm (defined as V_m), the number of vessel voxels in the prior annotated data (defined as V_l) and the number of vessel voxels in the refined annotated data (defined as V_n). It was found that the proportion of V_m to V_l (unlabeled vessel voxels ratio) in 20 3Dircadb datasets was between 1% and 55% with average value 16%. It can be seen, that the proportion of unlabeled voxels was still large and the quality of the annotations were quite high. Using the incomplete labelled annotations as reference for evaluation could lead to the evaluation bias.

The proportion of the number of over-segmented voxels by our algorithm to V_n was also studied to evaluate the performance of our method. It occupied nearly 8.8% of the annotated data. For most cases, the segmentation results were acceptable, and large errors occurred in liver data with extreme artifacts (such as #3 in 3Dircadb) or in data with large and extremely heterogeneous tumors (#17). Figs. 10 and 11 show the comparative result of original and refined annotated data with large or small unlabeled vessels. The 3D visualization was performed on the software of Amira.

As seen in Figs. 10 and 11, the unlabeled vessel voxels varied in different datasets. Using our proposed method, the original annotated data were refined as the new benchmark, and the extracted vessels were more complete, more continuous, and richer in the refined annotated data. The evaluation before was based on reference data with incomplete labeling and the assessment was biased. Re-evaluation is needed based on the refined annotated data. We have compared the evaluations based on the incomplete annotations (original expert manual segmentation results) and refined annotations without post-processing to determine the assessment bias on 20 3Dircadb datasets, as shown in Table 1.

In the re-evaluation, our average dice value was 75.3%, sensitivity was 76.7%, accuracy and specificity were 97.6% and 98.8% respectively. Compared with the evaluation before, the average dice value

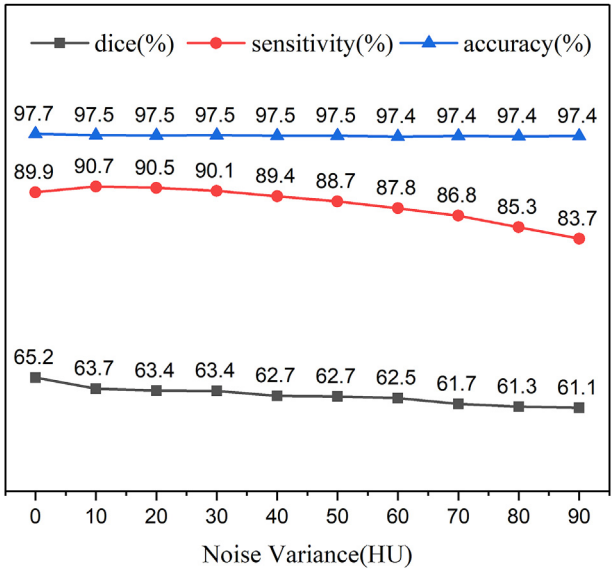

Fig. 8. Evaluation of the proposed method on synthetic data with different additive noise variances.

improved by 7.8% and sensitivity improved by 2.4%. When using the original annotated data as the testing data and the refined annotated data as the benchmark, the average dice value was 92.0% and sensitivity was 86.3%. As can be seen, certain deviations existed using incomplete annotated data and a large evaluation bias occurred.

3.6. 3D visualization of liver tumor and vessel

The relative 3D position relationship between the liver tumor and liver vessels is important in liver interventional surgery planning and guiding. Based on our liver vessel extraction and liver tumor segmentation method [31], we construct the 3D relationship of liver tumor and vessels. Fig. 12 and Fig. 13 show the 3D visualization of a small and a large liver tumor and liver vessels from the Sliver07 datasets and local datasets respectively.

In Fig. 12, the liver tumor was small and liver vessels around the tumor had been effectively and accurately extracted by our algorithm. Due to the effect of training samples (most inferior vena cava (IVC) near or attached to the liver were not classified as a liver part in 3Dircadb datasets), most of the IVC were not detected. In Fig. 13, the liver tumor was large and some vessels around the tumor were irregular. Small or large liver vessels in or near the tumor can still be effectively extracted from the dataset with high noise. Our proposed method provides an approach for constructing the 3D relative position of liver tumor and nearby liver vessels.

4. Discussion

For robust liver vessels extraction from CT images in liver surgery planning, we propose an automatic method using 3D U-Net with a variant dice loss function. The main contributions of the proposed algorithm are as follows: (1) Using the deep-learning method for accurate and automatic liver vessel extraction with few training samples and incomplete annotations. (2) Proposing a new loss function to deal with unbalanced classes and improve segmentation accuracy and sensitivity, with better results than the network with dice loss function. (3) Providing a robust liver vessel extraction method in images with high noise and liver tumors. (4) Showing the impact of benchmark quality on evaluations to remind the researcher that supervised learning methods should further consider quality of the annotation and the real results, to prevent evaluation bias and to improve the medical analytical result.

While deep convolutional networks can be effectively used for image segmentation, they still heavily rely upon the number and types of training samples. For medical image datasets, the datasets are much smaller than natural image datasets, and the training samples are limited. Manual annotation of medical datasets with complex structures were time and energy consuming, and the results may be incomplete even for experts. Moreover, it is usually difficult for researchers to evaluate the professional skills of the annotated entries, and some of the benchmarks might even be incorrect. Besides, Unbalanced classes of the foreground and background can also cause the segmentation error.

On one aspect, the choice of network and loss function affects the performance of deep learning-based segmentation methods greatly. For accurate and robust liver vessel extraction from CT images with few training samples and incomplete labeling, data augmentation is used to enrich the appearances of liver vessels and improve the robustness of the training network, and 3D U-Net dense network is selected for 3D image feature extraction with sparse annotation. To improve segmentation accuracy with imbalanced classes, we adjust the parameters of the number of correctly classified foreground voxels and the number of misclassified voxels based on the dice loss function, and increase the penalty for misclassified voxels to teach the network to identify vessels with weak boundary, high noise or low contrast. As shown in Figs. 4–6, the vessels with fuzzy boundaries, high noise or heterogeneous densities were more effectively and accurately identified by our method than method with dice loss function, and accuracy and sensitivity had improved using our method.

We found the potential of the deep-learning method to extract unannotated voxels that should be segmented, and refined the public annotated dataset (3Dircadb datasets) by including these voxels, considering the completeness of the annotation. We then compared the evaluation bias using the original and refined annotated dataset as benchmarks. On average nearly 16% of the liver vessels are unlabeled, and training or evaluation with incomplete annotation is biased. Current segmentation methods often used annotated datasets as the benchmark for training or evaluation without considering the quality of the annotation. Improving the supervised method for improved result (including evaluation scoring) is important; researchers should also

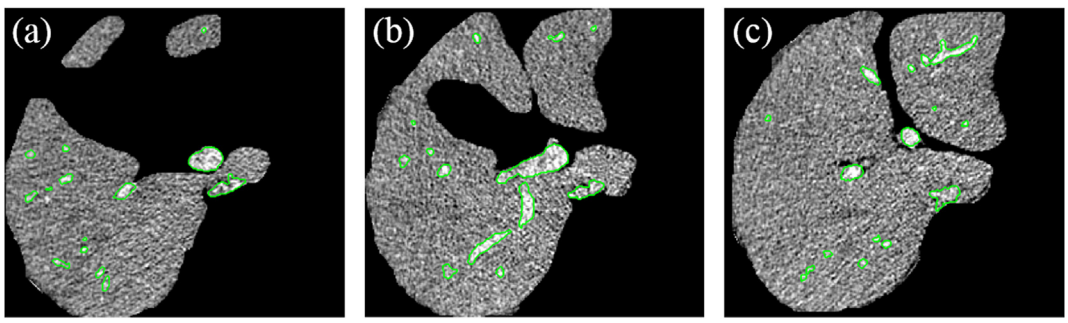

Fig. 9. Examples of performances of the proposed method on slices with separated liver partitions.

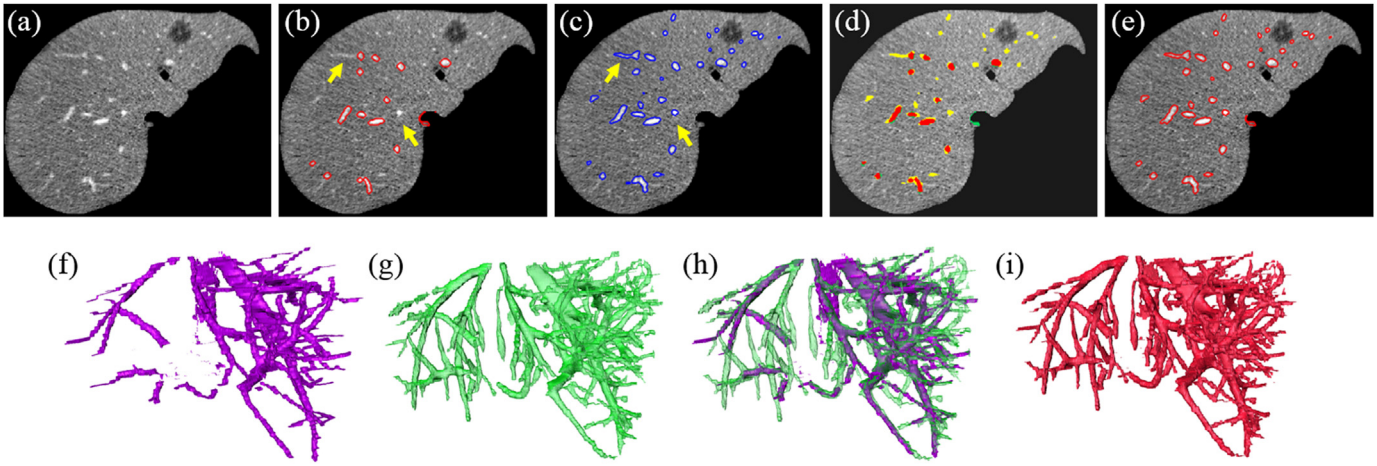

Fig. 10. Original annotated data ((b) (f)) with large unlabeled vessels and refined annotated data ((e) (i)). (a) CT slice; (c) (g) performance of our algorithm; (d) (h) fusion of performance by expert and our algorithm. Red for overlap area, yellow and green are for over-segmented or under-segmented area compared to the expert result. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

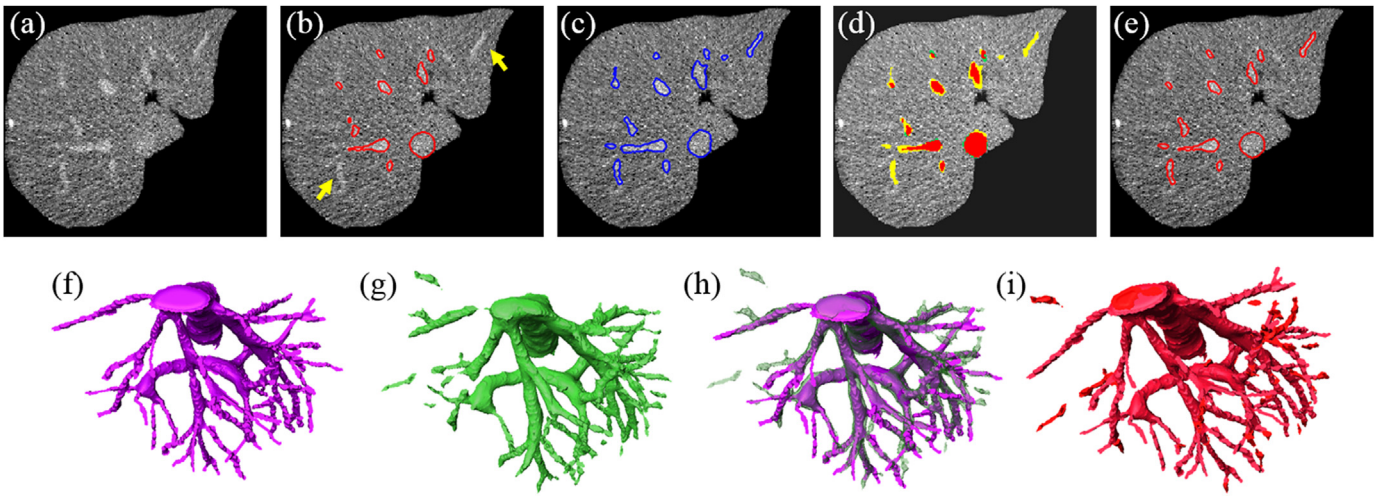

Fig. 11. Original annotated data ((b) (f)) with small unlabeled vessels and refined annotated data ((e) (i)). (a) CT slice; (c) (g) performance of our algorithm; (d) (h) fusion of performance by expert and our algorithm. Red for overlap area, yellow and green are for over-segmented or under-segmented area compared to the expert result. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Comparative Evaluation with original and refined annotated data.

Testing	Reference	Dice (%)	Sensitivity (%)	Accuracy (%)	Specificity (%)
Our	Original	67.5 ± 6.9	74.3 ± 10.6	97.1 ± 0.8	98.3 ± 0.8
Our	Refined	75.3 ± 4.4	76.7 ± 11.2	97.6 ± 0.8	98.8 ± 0.6
Original	Refined	92.0 ± 6.8	86.3 ± 10.4	99.4 ± 0.3	100.0 ± 0.1

consider the quality of the annotations and whether the results are similar to the real results. Our evaluation result was much closer to the clinical results, rather than results used for comparison based on incomplete annotation.

It has been proved that the proposed algorithm is accurate and robust in liver vessels extraction from CT images with high noise (see Figs. 7, Figs. 11 and 13), different liver vessel structures (see Figs. 6 and 11, tubular and elliptical and some irregular structures), varied intensity distributions (see Fig. 6) and separated liver partitions (see Fig. 9). The average dice value, sensitivity and accuracy of the proposed method on 20 3Dircadb datasets were 75.3%, 76.7% and 97.6% respectively. Kitrungrotsakul et al. [20] reported the average dice value of 7 datasets was 83%, while in their results unlabeled vessels were not extracted and the algorithm was sensitivity to intensity change. While the accuracy of traditional hessian-based methods was up to 85% [10], the average overlap error and sensitivity of context-based voting were 55% and 70% respectively [3]. The proposed method can effectively achieve liver vessel extraction and construct the relative 3D position relationship of liver tumors and vessels (see Figs. 12 and 13). It can replace the manual segmentation of liver vessels in clinical settings and be used for rough annotation of new datasets.

There are also some limitations. First, limited by the number and type of training samples, some liver vessels are not detected (see Fig. 12). Second, some background voxels are misclassified in livers with large and extreme heterogeneous tumors (high iodipin in tumor). In the future, we will collect more training samples to improve the robustness of the network and evaluate our method with more datasets. The refined annotated datasets will be used for re-training to see its

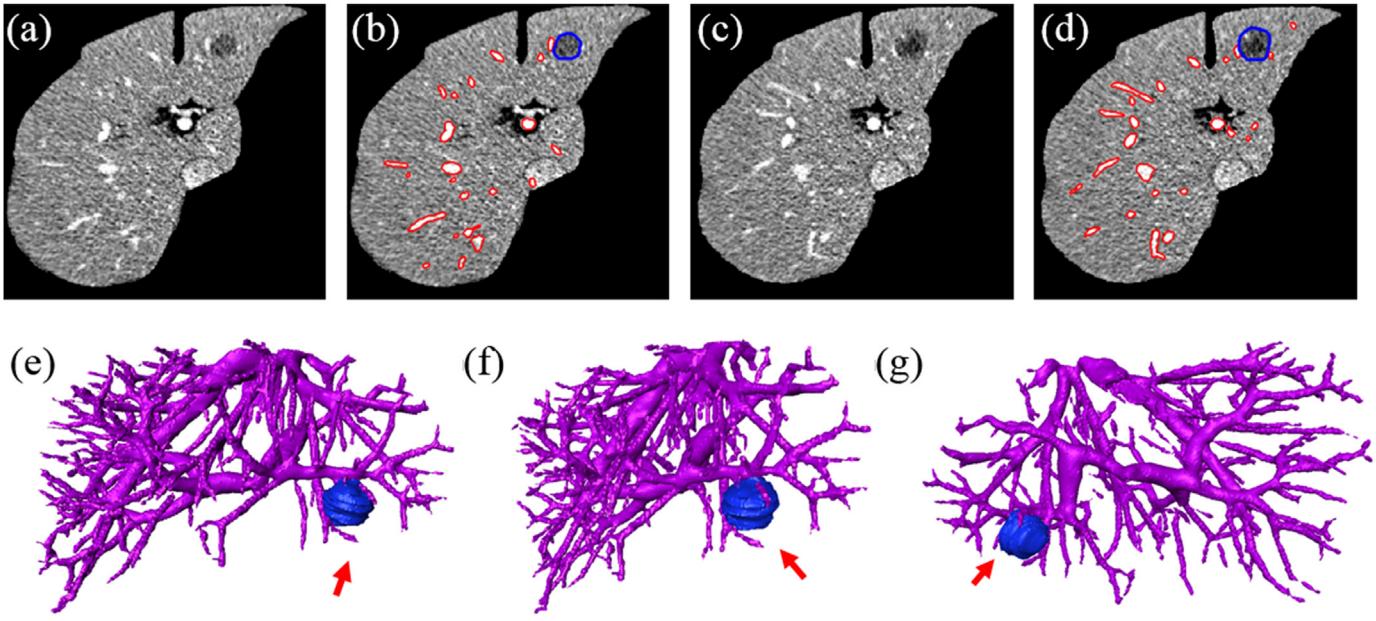

Fig. 12. Visualization of a small liver tumor and nearby liver vessels on Sliver07 datasets. (a) (c) CT slices; (b) (d) results of liver tumor (blue line) and liver vessels (red line); (e) (f) (g) 3D visualization of different views. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

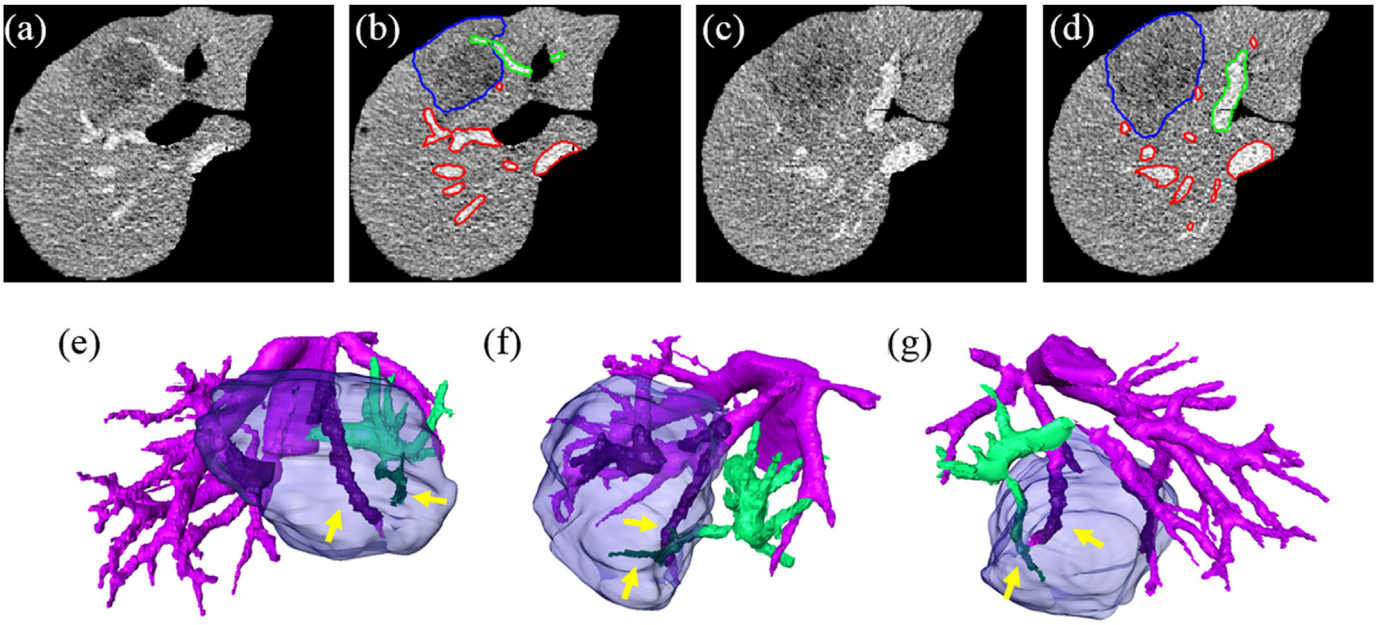

Fig. 13. Visualization of a large liver tumor and nearby liver vessels on local datasets. (a) (c) CT slices; (b) (d) results of liver tumor (blue line) and liver vessels (red line); (e) (f) (g) 3D visualization of different views. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

potential for improving segmentation performance, and the relationship between liver tumor and liver vessels will be further studied.

5. Conclusions

This paper presents an automatic liver vessels extraction method from CT images using a fully convolutional network. 3D U-Net dense network is chosen and data augmentation is used for training with few training samples and incomplete annotation. A new similarity metric based on the dice coefficient is proposed and used for the loss function to improve segmentation accuracy and sensitivity with unbalanced foreground and background class voxels. The public annotated datasets are refined by including unlabeled vessels that are extracted by our method and used for algorithm re-evaluation. Evaluation bias based on different benchmarks was calculated to show the impact of benchmark quality for a supervised learning method. Our method has been tested on 20 Sliver07 datasets, 20 3Dircadb datasets and local clinical datasets. The algorithm can realize the accurate and robust extraction of liver vessels from CT images with high noise, varied intensity distributions and liver structures effectively, and is superior to the method using the dice loss function. It can be used to replace the manual segmentation of liver vessels in clinic, build the 3D relationship of liver tumor and liver vessels and for rough annotation of new datasets.

Conflicts of interest

None declared.

Acknowledgments

This work was supported by the National Natural Science Foundation of China [grant numbers 81471759]. We appreciated Julia Wu from MIT for proof reading and correction of the manuscript.

References

[1] C. Schumann, J. Bieberstein, S. Braunewell, M. Niethammer, H.-O. Peitgen, Visualization support for the planning of hepatic needle placement, Int. J. Comput. Assist. Radiol. Surg. 7 (2012) 191–197.
[2] H.-W. Huang, Influence of blood vessel on the thermal lesion formation during radiofrequency ablation for liver tumors, Med. Phys. 40 (2013) 073303-n/a.
[3] Y. Chi, J. Liu, S.K. Venkatesh, S. Huang, J. Zhou, Q. Tian, W.L. Nowinski, Segmentation of liver vasculature from contrast enhanced CT images using context-based voting, IEEE (Inst. Electr. Electron. Eng.) Trans. Biomed. Eng. 58 (2011) 2144–2153.
[4] S. Moccia, E. De Momi, S. El Hadji, L.S. Mattos, Blood vessel segmentation algorithms — review of methods, datasets and evaluation metrics, Comput. Meth. Progr. Biomed. 158 (2018) 71–91.
[5] A.H. Foruzan, R.A. Zoroofi, Y. Sato, M. Hori, A Hessian-based filter for vascular segmentation of noisy hepatic CT scans, Int. J. Comput. Assist. Radiol. Surg. 7 (2012) 199–205.
[6] Q. Shang, L. Clements, R.L. Galloway, W.C. Chapman, B.M. Dawant, Adaptive directional region growing segmentation of the hepatic vasculature, Medical Imaging, SPIE, 2008, p. 10.
[7] V. Pamulapati, B.J. Wood, M.G. Linguraru, Intra-hepatic vessel segmentation and classification in multi-phase CT using optimized graph cuts, IEEE International Symposium on Biomedical Imaging, From Nano to Macro, 2011, pp. 1982–1985 2011.
[8] G. Läthén, J. Jonasson, M. Borga, Blood vessel segmentation using multi-scale quadrature filtering, Pattern Recogn. Lett. 31 (2010) 762–767.
[9] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen, S.A. Barman, Blood vessel segmentation methodologies in retinal images – a survey, Comput. Meth. Progr. Biomed. 108 (2012) 407–433.
[10] L. Ha Manh, K. Camiel, M. Adriaan, N. Wiro, W. Theo van, Quantitative evaluation of noise reduction and vesselness filters for liver vessel segmentation on abdominal CTA images, Phys. Med. Biol. 60 (2015) 3905.
[11] K. Drechsler, C.O. Laura, Comparison of vesselness functions for multiscale analysis of the liver vasculature, Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, 2010, pp. 1–5.
[12] Y. Tian, Q. Chen, W. Wang, Y. Peng, Q. Wang, F. Duan, Z. Wu, M. Zhou, A vessel active contour model for vascular segmentation, BioMed Res. Int. (2014) 15 2014.
[13] M. Freiman, L. Joskowicz, J. Sosna, A Variational Method for Vessels Segmentation: Algorithm and Application to Liver Vessels Visualization, SPIE Medical Imaging, SPIE, 2009, p. 8.
[14] O. Friman, M. Hindennach, C. Kühnel, H.-O. Peitgen, Multiple hypothesis template tracking of small 3D vessel structures, Med. Image Anal. 14 (2010) 160–171.
[15] S. Cetin, A. Demir, A. Yezzi, M. Degertekin, G. Unal, Vessel tractography using an intensity based tensor model with branch detection, IEEE Trans. Med. Imag. 32 (2013) 348–363.
[16] S. Cetin, G. Unal, A higher-order tensor vessel tractography for segmentation of vascular structures, IEEE Trans. Med. Imag. 34 (2015) 2172–2185.
[17] F. Bukenya, A. Kiweewa, L. Bai, A Review of Vessel Segmentation Techniques, (2017).
[18] E. Goceri, Z.K. Shah, M.N. Gurcan, Vessel segmentation from abdominal magnetic resonance images: adaptive and reconstructive approach, Int. J. Numer. Meth. Biomed. Eng. 33 (2017) e2811.
[19] Y.Z. Zeng, Y.Q. Zhao, M. Liao, B.J. Zou, X.F. Wang, W. Wang, Liver vessel segmentation based on extreme learning machine, Phys. Med. 32 (2016) 709–716.
[20] T. Kitrungrotsakul, X.-H. Han, Y. Iwamoto, A.H. Foruzan, L. Lin, Y.-W. Chen, Robust Hepatic Vessel Segmentation Using Multi Deep Convolution Network, SPIE Medical Imaging, International Society for Optics and Photonics, 2017 1013711-1013711-1013716.
[21] Ö. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: S. Ourselin, L. Joskowicz, M.R. Sabuncu, G. Unal, W. Wells (Eds.), Medical Image Computing and Computer-assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II, Springer International Publishing, Cham, 2016, pp. 424–432.
[22] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (Eds.), Medical Image Computing and Computer-assisted Intervention – MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III, Springer International Publishing, Cham, 2015, pp. 234–241.
[23] F. Milletari, N. Navab, S.A. Ahmadi, V-Net, Fully convolutional neural networks for volumetric medical image segmentation, Fourth International Conference on 3D Vision (3DV), 2016, 2016, pp. 565–571.
[24] S.S.M. Salehi, D. Erdogmus, A. Gholipour, Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks, Springer International Publishing, Cham, 2017, pp. 379–387.
[25] A. Nikonorov, A. Kolsanov, M. Petrov, Y. Yuzifovich, E. Prilepin, S. Chaplygin, P. Zelter, K. Bychenkov, Vessel Segmentation for Noisy CT Data with Quality Measure Based on Single-Point Contrast-to-noise Ratio, Springer International Publishing, Cham, 2016, pp. 490–507.
[26] G. Pizaine, E.D. Angelini, I. Bloch, S. Makram-Ebeid, Vessel geometry modeling and segmentation using convolution surfaces and an implicit medial axis, IEEE International Symposium on Biomedical Imaging, From Nano to Macro, 2011, pp. 1421–1424 2011.
[27] Q. Huang, H. Ding, X. Wang, G. Wang, Fully automatic liver segmentation in CT images using modified graph cuts and feature detection, Comput. Biol. Med. 95 (2018) 198–208.
[28] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: B. Francis, B. David (Eds.), Proceedings of the 32nd International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, 2015, pp. 448–456.
[29] D.P. Kingma, J. Ba, Adam, A Method for Stochastic Optimization, (2014) arXiv preprint arXiv:1412.6980.
[30] A.M. Mendonca, A. Campilho, Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction, IEEE Trans. Med. Imag. 25 (2006) 1200–1213.
[31] Q. Huang, H. Ding, X. Wang, G. Wang, Robust extraction for low-contrast liver tumors using modified adaptive likelihood estimation, Int. J. Comput. Assist. Radiol. Surg. (2018).