Correlation Matters: Multi-scale Fine-Grained Contextual Information Extraction for Hepatic Tumor Segmentation

Shuchao Pang[1(✉)], Anan Du[2], Zhenmei Yu[3], and Mehmet A. Orgun[1]

[1] Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
pangshuchao1212@sina.com, mehmet.orgun@mq.edu.au
[2] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia
anan.du@student.uts.edu.au
[3] School of Data and Computer Science, Shandong Women's University, Jinan 250014, China
zhenmei_yu@sdwu.edu.cn

Abstract. Automatic tumor segmentation has been used as a diagnostic aid in the identification of diseases such as tumors from liver CT scans, and their treatment. Owing to their success in computer vision tasks, the state-of-the-art Fully Convolutional Networks (FCNs) or U-Net based models have often been employed in many recent studies for automatic tumor segmentation to learn numerous weight-shared convolutional kernels and extract various semantic features. However, the correlation between different tumor regions in feature maps cannot be easily captured due to the lack of contextual dependencies, which in turn limits the representative capability of the adopted models and thus affects the accuracy of tumor segmentation results. To resolve this issue, we propose a novel framework for segmentation of tumors in liver CT scans, which can explicitly extract multi-scale fine-grained contextual information by adaptively aggregating local features with their global dependencies. The proposed multi-scale framework features a light model with a very few additional parameters, and also its visualization capability significantly boosts networks' interpretability. Experimental results on a real-world liver tumor CT dataset illustrate that the proposed framework achieves the state-of-the-art performance in terms of a number of widely used evaluation criteria for the hepatic tumor segmentation task.

Keywords: Hepatic tumor segmentation · Contextual information · Visualization · FCNs

1 Introduction

According to the latest liver cancer statistics from World Cancer Research Fund International and American Institute for Cancer Research, liver cancer was the sixth most common cancer worldwide in 2018 [1]. In particular, it is the ninth most commonly occurring cancer in women, but the fifth most common cancer in men. Furthermore, there were more than a total of 840,000 new cases diagnosed in 2018 which was 1.074

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 462–474, 2020.
https://doi.org/10.1007/978-3-030-47426-3_36

times more than that in 2012 [2, 3]. Besides having a healthy diet and being physically active, an early detection and intervention is also critical in mitigating the risk of liver cancer. Currently, with the rapid development of medical imaging technology, CT and MRI medical imaging examinations have been widely used in clinical applications to monitor the liver structure and state for diagnosis and treatment of liver cancer [4]. However, manually analyzing detected imaging slices is really a time-consuming and error-prone task to conduct for physicians and radiologists alike and there often exist some inter-observer variations for this kind of pixel-level labelling tasks [5]. Therefore, an accurate and automatic hepatic lesions/tumors localization and segmentation approach is urgently required as a diagnostic aid for early liver cancer detection.

However, in medical tumor segmentation tasks from liver scans, there still exist several hard challenges, with hepatic tumors as an example, such as low tissue contrast, large variability in tumor shape, size and number among inter-patient CT scans and intra-patient slices, and the vague boundary problem between diseased and healthy regions in the whole liver. In recent years, Fully Convolutional Networks (FCNs) [6] and U-Net [7] based deep neural networks have been widely utilized in biomedical and medical image segmentation tasks with an outstanding success [9, 11, 12]. Both types of network architectures utilize skip connections to integrate shallow feature maps and high semantic feature maps from different scales, which can generate more precise pixel-level recognition by fusing detailed positional information from shallow layers. Still, it should be noted that the range of contextual information obtained from those models is heavily limited by the depth of networks and the size of kernels used. Several recent works [13] modify these basic architectures by introducing multi-scale context fusion motivated by the Inception-ResNet-V2 model, where a large reception field can extract more abstract features for large objects, while a small reception field is better for small objects. Even though fusing multi-scale contextual information can capture different size objects, it cannot leverage the correlation between different objects in a global context, which is very important for medical tumor segmentation, in particular, segmenting common multiple tumors in a liver. To further exploit contextual dependencies, U-Net variants based on Recurrent Neutral Networks (RNNs) have been proposed to aggregate the context over local features from output feature maps of top layers of pre-trained CNN models [16]. Despite the enhancement of their representative capability, the implicitly captured global dependencies heavily rely on the learning outcome of the long-term memorization [17].

Different from these contextual extraction modules, in this paper, we propose a multi-scale contextual dependency framework inspired by attention mechanisms in machine translation tasks [14] to capture fine-grained contexts for inter- and intra-tumor regions and enhance the discriminability of learned features, and thus improve the performance in the hepatic tumor segmentation task, as shown in Fig. 1. More specifically, we first construct a new U-shape model motivated by CE-Net [13], where the pre-trained ResNet model and different size context aggregation with dilated convolutions and a multi-kernel pyramid pooling are fused into an encoder-decoder architecture. Then, we place the multi-scale context extraction model on all the skip connections to capture fine-grained contextual information by adaptively aggregating local features with their global dependencies from different scale feature maps, respectively. Finally, for a context extraction

block on each skip connection, we model the semantic context interdependencies over all the local features from both the spatial and the channel dimensions. In this way, the spatial contextual relationship can avoid the effect of the position distance between tumor regions in 2D feature maps and meanwhile, aggregate tumor features at each location by summing a global dependency on all the related tumor features. Furthermore, a global interdependent channel affinity map is also computed to exploit and emphasize the correlation among different feature categories along the channel dimensionality. By adding the two-level extracted contextual information element-by-element, the explicit fine-grained contexts can be learnt to produce more precise predictions for hepatic tumor segmentation, especially for small tumors. Moreover, with the guidance of the learned multi-scale contextual dependencies, the false-positive results are also significantly reduced, which is quite important for early cancer detection due to the existence of small lesions or tumor regions in the early stages. Furthermore, the interpretability of the proposed networks has also been greatly improved for hepatic tumor segmentation.

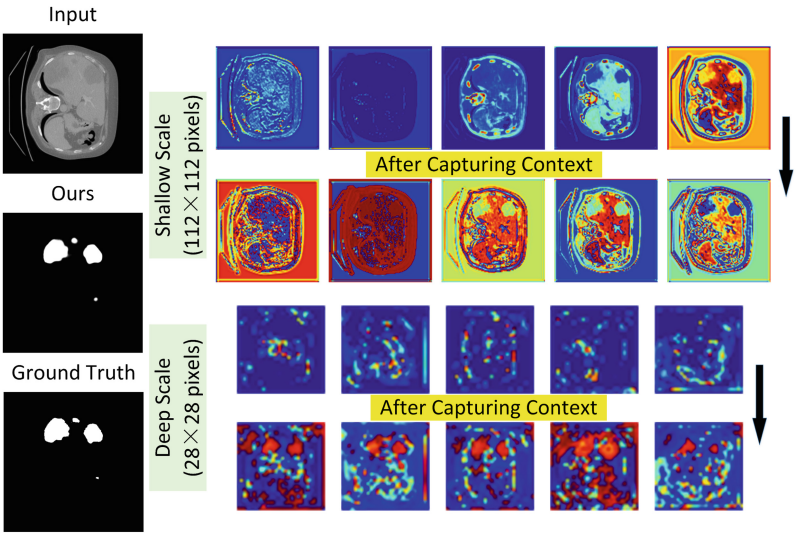

Fig. 1. A test example with our segmentation result and internal learned feature visualization comparison before and after using our multi-scale contextual dependency framework. Several feature map pairs corresponding to two scale contextual operations are respectively given and each learned feature map is enlarged for clarity. Note that the width and height of input image is set to 448 × 448 pixels in our networks.

Contributions of this study can be summarized as follows:

- We propose a novel framework to explicitly aggregate contextual relationships between hepatic tumors in different scale feature maps, which can successfully address various complex and hard challenges in medical tumor segmentation. The proposed framework is also an important improvement over the current automatic segmentation

methods. Moreover, parts or all of the proposed framework can be integrated into any FCNs or U-Net based architectures seamlessly.

- Our proposed global dependency extraction module operates on all skip connections to capture multi-scale fine-grained hepatic tumor contextual information, where two types of context aggregations are embedded into each skip connection for exploiting long-range contextual dependencies from both tumor spatial and channel dimensionalities. In addition, the explicit context aggregation with feature visualization noticeably boosts model's interpretability.
- The proposed medical tumor segmentation framework has been evaluated on real-world hepatic tumor data. The results show that multi-scale contextual dependencies over feature spatial regions and channel maps have significantly improved tumor segmentation performance, while reducing false positive and false negative rates of hepatic tumors on CT slices, and they have also enhanced the discriminative ability of learned representations in medical tumor segmentation.

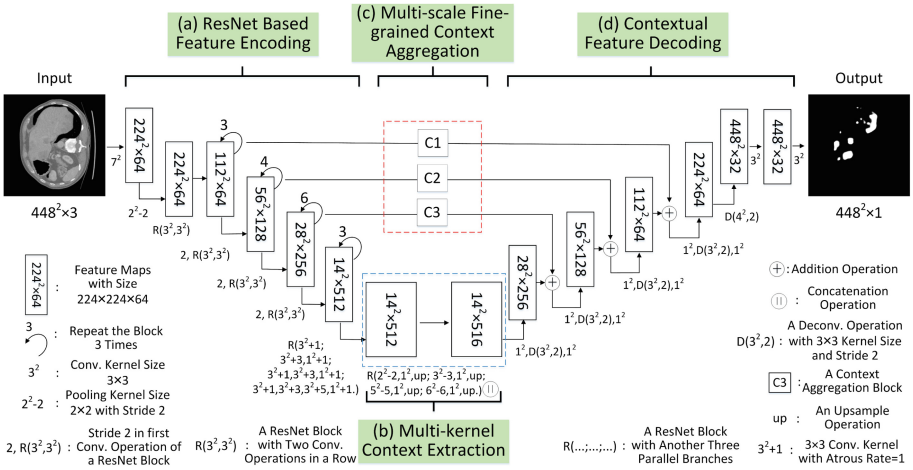

Fig. 2. Our multi-scale fine-grained contextual dependency framework for hepatic tumor segmentation, which consists of several main functional modules: (a) a ResNet-34 based feature encoding module, (b) a multi-kernel context extraction module, (c) a multi-scale fine-grained contextual aggregation module and (d) a contextual feature decoding module. An example with its prediction from the proposed algorithm is illustrated end-to-end in the whole workflow.

2 The Proposed Multi-scale Framework

2.1 Overview

In this paper, we propose a multi-scale fine-grained contextual information extraction framework to model long-range contextual dependencies over CT imaging regions for improving hepatic tumor segmentation performance. The proposed network framework

can perform global context aggregation over locally connected feature maps and then embed their global dependencies into local features, which can further increase the correlations between tumor regions and enhance their representative capability for medical tumor segmentation. By explicitly passing similar local contexts regardless of positional distances like in an undirected graph operation, the correlation and interaction of contextual dependencies from both the spatial and the channel dimensions is explicitly propagated and encoded into subsequent feature maps. Moreover, the characteristics of small-size tumor/lesion regions can also be inferred better after they are perfectly contextualized by utilizing this multi-scale contextual design, which noticeably reduces false positive cases as well as giving clear boundary predictions.

In order to take the full advantage of its effectiveness, the proposed multi-scale context framework is fused into a new U-shape context encoder network, which gives a significant improvement for the backbone and its variants, and really differentiates them in the aspect of context aggregation. Moreover, our proposed multi-scale framework requires a very few additional parameters, which only increases by 0.37% over that of the backbone networks. Experimental results show that our proposed multi-scale framework performs better than the state-of-the-art methods for medical tumor segmentation. The whole architecture of our designed networks is shown in Fig. 2, which includes four main parts: ResNet based feature encoding, multi-kernel context extraction, multi-scale fine-grained contextual aggregation and contextual feature decoding.

2.2 Spatial Context Extractor

There is a spatial context extractor for modeling the 2D contextual dependencies and a channel context extractor for modeling the 3D contextual dependencies on each of the three-dimensional feature map groups (where W × H × C refers to width, height and channel numbers of the learned features for each input image).

Subsequently, we introduce the spatial context extractor in detail and discuss the process of adaptively aggregating the 2D contextual dependencies. First, an input image with 448 × 448 × 3 size on the left in Fig. 2 is fed into the ResNet based feature encoding subnetwork for extracting its high-level semantic features. We assume that the learned 3D feature maps are a W × H × C tensor, where each 2D feature map is W × H pixels, the channel number is C and the batch size is set to 1 for clarity, like the input data $\boldsymbol{X} \in \mathbb{R}^{28\times28\times256}$ shown on the left side of Fig. 3. $V = \{v_i\}_{i=1:N}$ is the vertex set for each local contextual feature v_i at all 2D positions and N = W × H (also N = 784 in this example from Fig. 3). Then, in order to obtain the spatial contextual map among all the global spatial positions, two new feature maps $\boldsymbol{Q} \in \mathbb{R}^{28\times28\times32}$ and $\boldsymbol{K} \in \mathbb{R}^{28\times28\times32}$ are respectively generated with two single convolutional operations by 1 × 1 kernels, which is based on the fed feature map, as calculated by the following equations:

$$Q^{(v_i)} = f\left(w_1 X^{(v_i)} + b_1\right); K^{(v_i)} = f\left(w_2 X^{(v_i)} + b_2\right), \tag{1}$$

where these two operations can further encode each local positional context feature v_i and also reduce the parameters by reducing the channel dimensionality from 256 to 32, and f is a non-linear activation function and w_1, b_1, w_2, b_2 are network parameters. After reshaping them into $\boldsymbol{Q'} \in \mathbb{R}^{(28\cdot28)\times32}$ and $\boldsymbol{K'} \in \mathbb{R}^{(28\cdot28)\times32}$, we perform a 2D

matrix multiplication between the $\boldsymbol{K'}$ and the transposed $\boldsymbol{Q^t}$, which aims to calculate the mutual similarity of any two local contextual features $v_i \in \boldsymbol{K'}$ and $v_j \in \boldsymbol{Q^t}$. In this way, the spatial context map $\boldsymbol{M}$ with global knowledge is generated, which represents the interdependency of local features from any two positions in a 2D spatial context. By applying a softmax operation to it as shown below, the updated context map M_{ij} can indicate a greater correlation between the two positions if their similarity value is larger.

$$M_{ij} = e^{v_i \cdot v_j} / \sum_{j=1}^{N} e^{v_i \cdot v_j}. \tag{2}$$

Later, for aggregating all positional local contextual information with global context dependencies for fine-grained spatial context extraction, another new feature map $\boldsymbol{F'} \in \mathbb{R}^{(28 \cdot 28) \times 256}$ is also produced by performing a convolutional layer on the original fed feature map $\boldsymbol{X}$ without a channel dimensionality reduction and a reshaped operation successively. After that, a context aggregation operation is performed to generate the aggregated feature map $\boldsymbol{X'} \in \mathbb{R}^{(28 \cdot 28) \times 256}$ by a matrix multiplication operation between $\boldsymbol{M}$ and $\boldsymbol{F'}$, where each position in $\boldsymbol{X'}$ represents its corresponding weighted summarization of features across all the positions. The aggregated feature map $\boldsymbol{X'}$ is then reshaped into a new $\boldsymbol{X'} \in \mathbb{R}^{28 \times 28 \times 256}$. Finally, local contextual features at each position from the original input feature map $\boldsymbol{X}$ are fused with their global contextual dependencies $\boldsymbol{X'}$ by an addition operation as in the following equation.

$$\boldsymbol{X^s} = \alpha \boldsymbol{X'} + \boldsymbol{X} \tag{3}$$

where $\boldsymbol{X^s}$ is the selectively aggregated contextual features by fusing local contexts and global contexts and α is a learnable scale parameter. Overall, the spatial context extraction as shown in Fig. 3(1) is completed in the whole 2D spatial positions, where the fine-grained context features can further improve intra-class compact and semantic consistency and contribute to enhancing hepatic tumor segmentation performance.

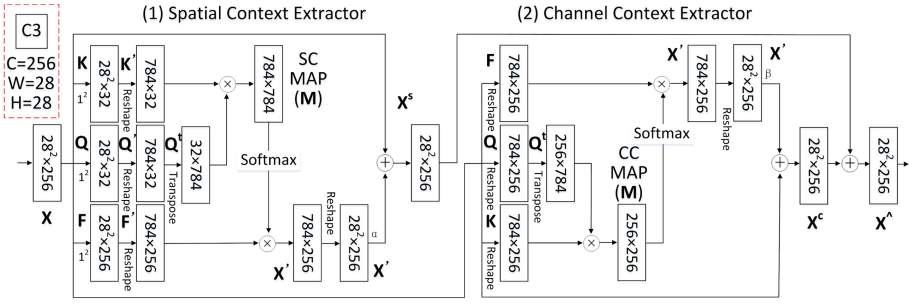

Fig. 3. A fine-grained contextual information aggregation block, taking C3 in Fig. 2 as an example. The details of the spatial context extractor and those of the channel context extractor for capturing rich contextual dependencies are illustrated in (1) and (2), respectively. Note that the operation $\otimes$ indicates matrix multiplication.

2.3 Channel Context Extractor

We observe that above context process just considers the 2D spatial positions by leveraging each local context v_i where v_i is a C dimensionality vector, which means that the interdependency and the correlation between different channels is not fully exploited. However, the 3D channel context information is essential to extract robust hepatic tumor knowledge. Therefore, this subsection discusses the extraction of channel contextual information. Different channel feature maps usually represent different image feature types and semantic information. Furthermore, semantic information from different channels are usually associated with each other, which can improve the representative capability of feature maps if we exploit them in global knowledge. So, in order to explicitly model the interdependencies between the channel maps, we respectively build a channel context extractor for each contextual information aggregation block from our proposed multi-scale fine-grained context extraction framework.

As illustrated in Fig. 3(2), a light channel context model is utilized to achieve fewer parameters in the process. When the input original feature maps $\boldsymbol{X} \in \mathbb{R}^{28\times28\times256}$ are fed into the spatial context extractor, we also deliver them into the channel context extractor in the meantime. Different from the former step, $\boldsymbol{X}$ is directly reshaped into $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{F} \in \mathbb{R}^{(28\cdot28)\times256}$ without any convolutional operation. Besides, $V = \{v_i\}_{i=1:C}$ is the channel set for each channel contextual feature $v_i \in \mathbb{R}^{28\cdot28}$ at the third dimension. Then, we perform a matrix multiplication between the transposed $\boldsymbol{Q^t}$ and $\boldsymbol{K}$ to calculate the channel similarity of any two channel maps over all the spatial positions. Later, the generated channel context map $\boldsymbol{M}$ is applied by a softmax layer to normalize them for satisfying the properties of probability. In addition, the global contextual dependency extraction $\boldsymbol{X'}$ for each channel map is obtained by a matrix multiplication along the channel dimension. To this end, the following equations are used.

$$\boldsymbol{M} = \boldsymbol{Q^t K}; M_{ij} = e^{v_i \cdot v_j} / \sum_{j=1}^{C} e^{v_i \cdot v_j}, v_i \in \boldsymbol{Q^t}, v_j \in \boldsymbol{K}. \tag{4}$$

$$\boldsymbol{X'} = \boldsymbol{FM}. \tag{5}$$

Finally, the obtained global contextual aggregation result $\boldsymbol{X'}$ with a parameter β is added into each original channel feature map from $\boldsymbol{X}$ along the channel dimension.

$$\boldsymbol{X^c} = \beta \boldsymbol{X'} + \boldsymbol{X}. \tag{6}$$

Overall, the final feature of each channel $\boldsymbol{X^c}$ is constructed by fusing a weighted sum of all the channel feature maps and the original single feature map in each channel space, which successfully models the long-range context semantic dependencies among the channel maps to boost their representative ability for medical tumor segmentation. Based on an addition operation from both of these context extraction steps, each context aggregation block can fully exploit contextual information in a global view from the spatial and the channel perspectives, as shown in Fig. 3.

$$\boldsymbol{X^\wedge} = \boldsymbol{X^s} + \boldsymbol{X^c}. \tag{7}$$

3 Experiments and Analysis

3.1 Data and Implementation Details

Evaluation Dataset and Metrics. Tumor segmentation is a more difficult task than general body organ segmentation tasks due to vague boundaries between diseased and healthy tissues. For this task, a new and challenging hepatic tumor dataset [5] is used to show hepatic tumor segmentation performance for all the methods considered. This dataset consists of 131 abdominal 3D CT scans acquired from 131 subjects with different types of liver tumor diseases, e.g., primary tumor diseases and secondary liver tumors. These medical data were collected from clinical sites in the world with different CT scanners and acquisition protocols. Here, we can extract 7190 CT slices with tumor annotations. For comprehensive comparisons between different segmentation methods, a number of widely used evaluation metrics are utilized in our study, including Dice similarity coefficient, Hausdorff distance, Jaccard index, precision (also called positive predictive value), recall (also called sensitivity coefficient or true positive rate), specificity coefficient (also called true negative rate) and F1 score. Except Hausdorff distance, the others indicate that the larger results are better.

Parameter Setting. 2 NVIDIA CUDA cores with 4 logical GPUs, and 1 Intel Haswell E5-2670v3 CPU are used to train our proposed multi-scale segmentation framework. The batch size for each forward pass is 8 CT slices and the initial learning rate is set to 0.0002, which could be dynamically changed during the training process under the guidance of the variations of errors. If there is no reduction of errors in the next 10 epochs, then the learning rate would be cut in half. Meanwhile, we set the maximum training epoch to 400 with an early stopping strategy. When the generated error is no longer reduced in the next 20 epochs or the learning rate drops below 5e-7, the training process is finished. And the training data and test data are randomly split into 4:1 from all raw data with tumor annotations. In addition, some widely used data augmentation techniques are also used dynamically during our training process [13]. Only the basic and plain cross entropy loss is employed for better demonstrating the robustness of our model. The Adam optimizer is employed to optimize and update all the network parameters in our segmentation network. The compared state-of-the-art methods are also trained on our dataset according to their original papers.

Compared Methods. The nine state-of-the-art segmentation methods are chosen in our experiments based on several representative models, as baseline methods for comparison: (1) *U-Net Based Model*: U-Net [7] is a highly cited architecture; Attention UNet [9] adds a spatial attention scheme; Nested UNet [12] employs hot dense skip pathways. (2) *Context Based Model*: R2U-Net [11] utilizes recurrent and residual networks; CE-Net [13] embeds a multi-kernel context encoding mechanism like Inception architecture; Self-attention [8, 10] exploits spatial context information. And (3) *Attention Based Model*: SENet [15] uses channel attention mechanism; both DANet [17] and CS-Net [18] place self-attention schemes on the top of encoder stage, but with different network architectures. (4) *Fused Model*: Attention UNet [9] and Self-attention [8, 10].

3.2 Quantitative Analysis

For quantitative analysis, all the ten representative segmentation methods are evaluated on the test dataset, as shown in Table 1. The pioneering U-shape network, U-Net [7] can be used to predict hepatic tumor regions on CT slices but with an unsatisfactory performance (e.g., 73.62% in Dice) as well as its variant Nested UNet [12] (e.g., 73.58% in Dice), while the attention based variant Attention UNet [9] is better by around 5% than the original U-Net under different segmentation criteria. This is because Attention UNet can give a further refinement for learned features from the spatial dimension to highlight the salient features and suppress useless ones. Similarly, SENet [15] also improves the segmentation performance by embedding squeeze-and-excitation blocks after skip connections as a channel attention mechanism, in spite of a slight decline compared to Attention UNet. Then, we compare popular context based models for medical tumor segmentation. R2U-Net [11] only outperforms its backbone model (U-Net) by about 1% in segmentation accuracies by employing RNNs and residual connections to extract feature context features. By contrast, multi-kernel context encoder networks CE-Net [13] can achieve relatively better performance (such as 78.41% in Dice, 33.78 pixels in HD, 72.92% in Precision and 89.33% in Recall) with different evaluation measures like Attention UNet [9], where pretrained network parameters can also provide some help in improving segmentation performance together with multiple kernel contextual feature extraction.

Table 1. Comparison results of the state-of-the-art segmentation methods with widely used evaluation metrics for hepatic tumor segmentation. The numbers in bold represent the best results. Note that Hausdorff distance uses pixel units and others %.

Methods/Metrics	Dice	Hausdorff distance	Jaccard	Precision	Recall	Specificity	F1
U-Net [7]	73.62	52.65	63.67	67.46	86.70	99.76	75.88
Attention UNet [9]	78.70	37.13	69.32	72.93	89.65	99.83	80.43
R2U-Net [11]	74.55	46.04	64.46	68.38	87.27	99.77	76.68
Nested UNet [12]	73.58	46.89	63.55	67.39	86.95	99.76	75.93
CE-Net [13]	78.41	33.78	69.09	72.92	89.33	99.82	80.30
SENet [15]	77.88	39.09	68.55	72.39	89.16	99.82	79.90
Self-attention [8, 10]	76.49	38.78	66.80	70.82	88.72	99.79	78.76
DANet [17]	79.97	30.94	71.00	74.50	90.38	99.84	81.67
CS-Net [18]	78.90	32.90	69.45	73.03	90.03	99.83	80.64
Ours	**82.16**	**30.01**	**73.46**	**76.96**	**91.18**	**99.86**	**83.37**
Average Gain (↗)	*5.26*	*9.79*	*6.14*	*5.87*	*2.49*	*0.058*	*4.46*

For comparison with recent self-attention and non-local models [8, 10], we have integrated their original spatial context extraction module into our backbone networks

in lieu of ours. As we can see from Table 1, the Self-attention model [8, 10] drops two percentage points over multi-kernel context encoder networks [13]. Very recently, both DANet [17] and CS-Net [18] have exploited two types of self-attention models acting on the top of a feature encoder path from different pretrained network architectures with better performance than Attention UNet [9] and CE-Net [13], for example, 79.97% vs 78.90% in Dice, 30.94 pixels vs 32.90 pixels in HD, 74.50% vs 73.03% in Precision and 90.38% vs 90.03% in Recall. Moreover, these good segmentation results also illustrate that diverse self-attention strategies can further boost the feature representative capability of a model for accurate tumor localization and segmentation.

More importantly, our proposed multi-scale framework performs the best under all the evaluation metrics while outperforming nine compared state-of-the-art methods by an average of 5.26% in Dice coefficient, ranging from 2.19% to 8.58%. In terms of Jaccard index and Precision coefficient, our model also shows an average gain of 6.14% ranging from 2.46% to 9.91%, and 5.87% ranging from 2.46% to 9.57%, repectively. In addition, the true positive rate (TPR, also Recall coefficient) from our method is also significantly better with an average 2.49%. While Specificity coefficient with 0.058% increase, also called true negative rate (TNR), is also slightly better than all the other methods due to a small percentage of tumor regions in CT slices; F1 score of our model noticeably outperforms all the baseline methods by an average of 4.46% by leveraging Precision and Recall results. Last but not the least, our multi-scale context aggregation method exhibits an average reduction in Hausdorff distance of 9.79 pixels, which means that the boundaries from our segmentation results can better coincide with their corresponding ground truths from radiologists than those of the nine state-of-the-art methods. Overall, our proposed segmentation method can outperform those nine state-of-the-art segmentation methods because our multi-scale context guided information aggregation process can better encode global knowledge into local features with fine-grained representations from spatial and channel dimensions and other important modules in our networks also boost segmentation performance of our framework.

3.3 Qualitative Analysis

As shown in Fig. 4, several segmentation results from randomly chosen CT imaging slices are visualized to provide a qualitative comparison of different models. Both our proposed method and DANet [17] perform well on the first sample, but others falsely consider healthy regions as hepatic tumors. This is similar in the second sample, except that Attention UNet [9] also works well. However, from the third sample, we see that DANet [17] has difficulty to differentiate hepatic tumors from surrounding tissues. As a whole, in cases Fig. 4(3–5), the false positive rates of the state-of-the-art methods are really high, resulting in many mis-segmented regions. This would negatively affect an accurate diagnosis for patients with hepatopathy, especially for early stage patients. On the other hand, in the sixth sample, the compared models just give partial predictions for tumor regions with some undiagnosed cases, which means a high false negative rate from their models. More importantly, in some challenging cases (e.g., Fig. 4(7)), all the baseline methods completely fail. Overall, all these misdiagnoses generated from the state-of-the-art automatic segmentation methods could be due to a lack of effective context

extraction for accurate hepatic tumor segmentation. By contrast, our proposed multi-scale segmentation framework can extract fine-grained global context dependencies from spatial and channel dimensions and then aggregate them together with local features to generate more precise segmentation results, as depicted in Fig. 4.

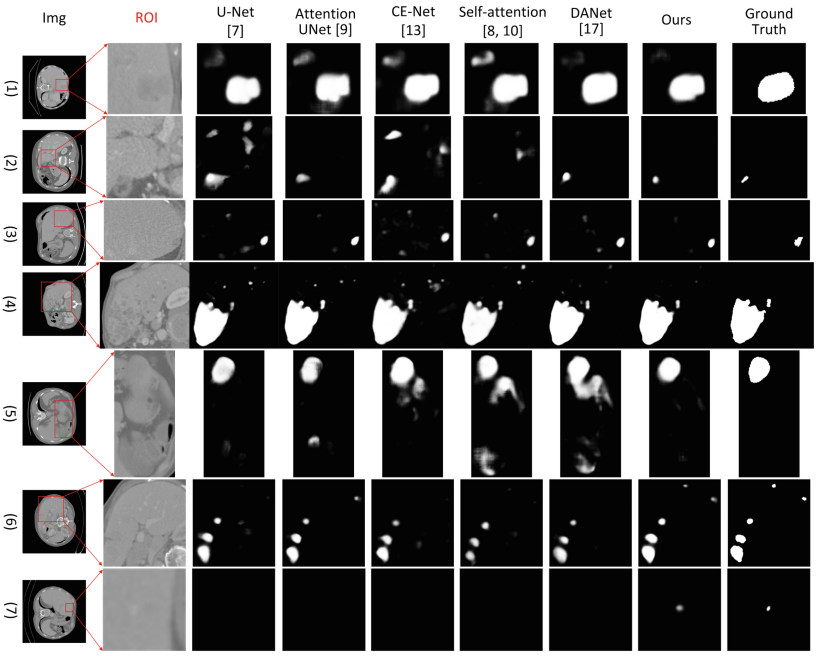

Fig. 4. Seven randomly selected samples with their segmentation results from the state-of-the-art methods. For clarity, we only report the regions of interest (ROI) of some of the compared methods due to space limitations.

4 Conclusions

In this paper, we have proposed a multi-scale contextual dependency framework to explicitly capture fine-grained context correlations between tumor regions and enhance the discriminability of the learned features and hence to improve segmentation performance for hepatic tumors. In particular, we have modeled the semantic context dependencies over all the local features from both the spatial and channel dimensions. To be specific, the spatial contextual relationship can aggregate tumor features at each spatial location by summing a global dependency on all related tumor features, which can lessen the effect of the position distance of local features in feature maps. On the other hand, a global interdependent channel affinity map is also computed to emphasize the correlation among different feature categories along the channel dimensionality. In addition, feature visualization analysis and comparison significantly improves the interpretability of

our proposed automatic segmentation networks. Extensive experiments conducted on a real-life liver tumor dataset also demonstrate that our model outperforms nine compared state-of-the-art segmentation methods. In the future, we plan to extend this framework into further clinical applications.

Acknowledgment. This work is supported in part by an International Macquarie University Research Excellence Scholarship (iMQRES: 2018150) and partially supported by the National Natural Science Foundation of China (No. 61472416).

References

1. Liver Cancer Statistics. https://www.wcrf.org/dietandcancer/cancer-trends/liver-cancer-statistics. Accessed 11 Dec 2019
2. Liver Cancer. https://www.wcrf.org/dietandcancer/liver-cancer. Accessed 11 Dec 2019
3. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. **68**(6), 394–424 (2018)
4. Budak, Ü., Guo, Y., Tanyildizi, E., Şengür, A.: Cascaded deep convolutional encoder-decoder neural networks for efficient liver tumor segmentation. Med. Hypotheses **134**, 109431 (2020)
5. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). arXiv preprint arXiv:1901.04056 (2019)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
7. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, Alejandro F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
8. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318 (2018)
9. Schlemper, J., et al.: Attention gated networks: learning to leverage salient regions in medical images. Med. Image Anal. **53**, 197–207 (2019)
10. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
11. Alom, M.Z., Yakopcic, C., Hasan, M., Taha, T.M., Asari, V.K.: Recurrent residual U-Net for medical image segmentation. J. Med. Imaging **6**(1), 014006 (2019)
12. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: UNet++: a nested U-Net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS-2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
13. Gu, Z., et al.: CE-Net: context encoder network for 2D medical image segmentation. IEEE Trans. Med. Imaging **38**, 2281–2292 (2019)
14. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
16. Shuai, B., Zuo, Z., Wang, B., Wang, G.: Scene segmentation with dag-recurrent neural networks. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1480–1493 (2017)

17. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
18. Mou, L., et al.: CS-Net: channel and spatial attention network for curvilinear structure segmentation. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11764, pp. 721–730. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32239-7_80