2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

HL-FCN: Hybrid Loss Guided FCN for Colorectal Cancer Segmentation

Yi-Jie Huang[1,2], *Qi Dou*[4], *Zi-Xian Wang*[3], *Li-Zhi Liu*[3], *Li-Sheng Wang*[1], *Hao Chen*[2,4⋆], *Pheng-Ann Heng*[4], *Rui-Hua Xu*[3⋆]

[1]Institute of Image Processing and Pattern Recognition, Department of Automation, Shanghai Jiao Tong University
[2]Imsight Medical Technology Co. Ltd.
[3]Sun Yat-sen University Cancer Center; State Key Laboratory of Oncology in South China;
Collaborative Innovation Center for Cancer Medicine, Guangzhou, China
[4]Department of Computer Science and Engineering, The Chinese University of Hong Kong

ABSTRACT

Colorectal cancer is among the leading cause of cancer-related mortalities. The cancerous regions are conventionally delineated from 3D magnetic resonance images in a voxel-wise way for radiotherapy. To ease the manual labeling procedure, which is laborious and time-consuming, automatic segmentation methods are highly demanded in clinical practice. However, it is a challenging task due to class imbalance and low-contrast appearance of cancerous regions, as well as the hard mimics from complex peritumoral areas. In this paper, we propose a volume-to-volume fully convolutional network architecture effectively trained with hybrid loss, referred as HL-FCN, to automatically segment colorectal cancer regions. Specifically, a novel Dice-based hybrid loss is designed under a multi-task learning framework to tackle the class-imbalance issue and hence improve the discrimination capability. Furthermore, a multi-resolution model ensemble strategy is developed to suppress false positives while preserving boundary details. Our method has been extensively validated on 64 cancerous cases using four-fold cross-validation, outperforming state-of-the-art methods by a significant margin.

***Index Terms*—** Multi-task FCN, hybrid dice loss, colorectal cancer segmentation, ensemble learning.

1. INTRODUCTION

Colorectal cancer is the second leading cause of cancer-related mortalities in the United States [1]. In current clinical routine, colorectal cancer regions are manually delineated from volumetric magnetic resonance (MR) images, which provide rich context of the soft tissue. However, this procedure is laborious, subjective and time-consuming, thus suffers from limited reproducibility. Therefore, automatic colorectal cancer segmentation methods are highly demanded in clinical practice.

However, this task is very challenging due to class imbalance, low-contrast appearance of cancerous regions, as well as hard mimics from the complex peritumoral background. As is shown in Fig. 1, it's hard to distinguish abnormality from the normal peritumoral tissue due to the small size, the ambiguity and inconsistency of appearance features (such as intensity distribution and shape), and the hard mimics from the complex peritumoral background.

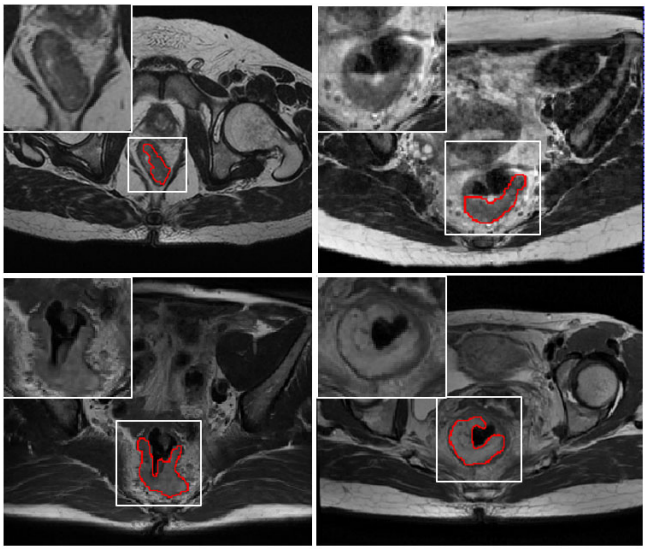

Fig. 1. Typical examples of MR slices with colorectal cancer. The cancer regions are delineated with red lines and zoomed in for clearer illustration.

Segmentation of MR images has been widely studied in the literature. For example, Mahapatra *et al.* proposed a low-level feature based super-voxel clustering manner to segment crohn disease from MR images [2]. Irving *et al.* further proposed a super-voxel-based clustering method to segment abnormal colon tissue [3]. Chen *et al.* [4] proposed a 3D fully convolutional networks for intervertebral disc localization and segmentation for MR images for the first time. Ronneberger *et al.* [5] proposed a 3D U-shaped architecture U-Net for medical image segmentation. Incorporated with skip

⋆ Corresponding authors.

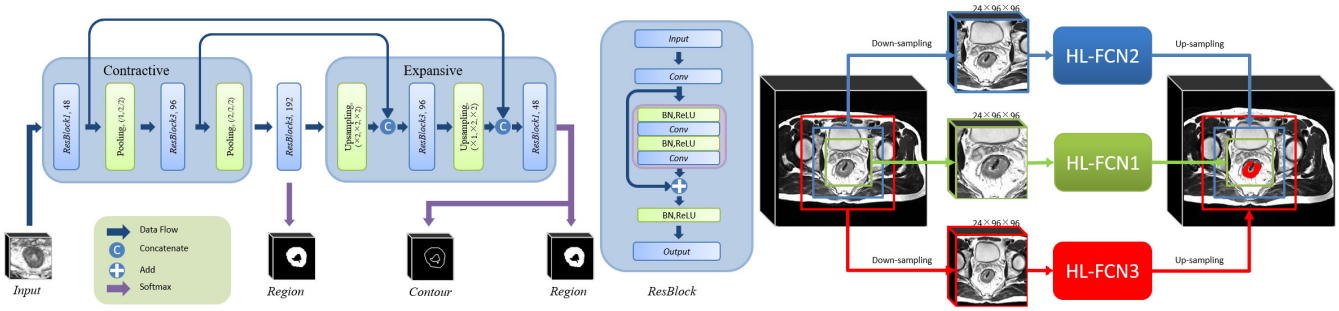

Fig. 2. Overview of the proposed HL-FCN architecture and the multi-resolution ensemble. ResBlock1 uses $1 \times 3 \times 3$ ZYX filter kernels and ResBlock3 uses $3 \times 3 \times 3$ filter ZYX kernels.

connections to encourage information propagation, volumetric FCN with residual connections was proposed to accelerate the convergence of deep learning and enhance the recognition capability [6, 7].

Different loss functions have been designed under the deep learning framework for semantic medical image segmentation. For example, Milletari *et al.* [8] adopted parameter-free Dice coefficient loss instead of cross-entropy to address class-imbalance issue. Chen *et al.* [9] proposed a contour-aware loss function for training to improve instance-wise segmentation ability. Dou *et al.* [10] employed deep supervision with additional supervision into hidden layers to accelerate training process and improve discrimination capability.

While three-dimensional FCNs serve good baselines in our underlying task by learning high-level features, there are several issues to be conquered and more effective learning strategies need to be explored. Firstly, class imbalance issue of learning for segmentation need to be resolved under the deep multi-task learning framework. Secondly, boundary details should be harnessed in an explicit way to enhance the segmentation performance. In addition, multi-resolution model ensemble can help to improve the segmentation results.

In this paper, we propose a volume-to-volume FCN architecture trained towards a hybrid loss named as HL-FCN to automatically segment colorectal cancer regions. Specifically, Dice coefficient is adopted to unify the loss formulation of different tasks, making the multi-task training procedure easier and the acquired models more effective. To further improve the performance, a multi-resolution model ensemble strategy is developed to suppress the false predictions and refine the boundary details. Cross-validation is conducted on the acquired dataset and extensive ablation studies demonstrate that each component of our method contributes to the performance gain.

2. METHODS

The workflow is shown in Fig. 2. Initially, images are resampled to three resolution rates. Subsequently, three volume-to-volume FCNs, namely HL-FCN1, HL-FCN2 and HL-FCN3, are trained and fused to generate final predictions.

2.1. Volume-to-volume 3D fully convolutional networks

The architecture of the proposed network is shown in Fig. 2. Since that the cancerous regions are volumetric data and cannot be discriminated from sole 2D context, the proposed model is designed to be a 3D end-to-end FCN. The network consists of a contractive path and an expansive path. The contractive path consists of 6 convolutional layers and 2 max pooling layers and performs cancerous region recognition. Subsequently, the expansive path further recovers shape details by employing 2 up-sampling layers, 6 convolutional layers and bridge concatenations to combine up-sampled high-level features with low-level features.

The input ZYX size is set as $24 \times 96 \times 96$ voxels, since the Z spacing of the acquired images is significantly larger than X and Y spacing. To tackle the anisotropic spacing, shallower layers employ flat kernels, referred as ResBlock1, to avoid neglecting details along the Z axis, leaving 3D context being handled by deeper layers.

To ease the training process, multiple auxiliary structures are added to the main architecture. Recently, residual learning shows effectiveness in accelerating convergence and improving performance [6], hence it is adopted between two pooling layers. In addition, deep supervision can ease the convergence and provide regularization to shallower layers for improving the discrimination capability [10]. Thus auxiliary loss functions with deep supervision are added into the overall loss.

In the test phase, overlapped sliding window is employed and the ZYX stride is set as (6,24,24). The overlapped predictions are averaged to generate the final output.

2.2. Dice-based multi-task hybrid loss function

It is observed that cancerous borders are hard to be learned due to the intensity ambiguity, previous studies validated the efficacy of contour exploration for semantic segmentation [9]. Therefore we added an auxiliary side task with contour-aware

boundary segmentation, trained in parallel with the region segmentation task. We added an extra softmax branch at the output to predict the contour voxels, as is shown in Fig. 2. The additional softmax layer shares the feature maps with the main softmax layer, thus is strongly related to the main prediction and helps the network learn more discriminative features.

Typically, voxel-wise cross-entropy loss is used for semantic segmentation and contour extraction [9]. In hard cases, mainly in border region, models trained towards cross-entropy loss tend to produce more uncertain response. The common practice to solve this issue is to assign extra weights to help the network focus more on the foreground, which in turn increases false positive rate. In contrast, Dice coefficient loss function, an automatic class balancer, is employed as a unified loss formulation to make training procedure easier. The Dice loss is defined as:

$$L_d = 1 - 2 \times \frac{\sum_{i=1}^{N} p_i g_i + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{1}$$

where the sums are computed over the N voxels of the predicted volume $p_i \in P$ and the ground truth volume $g_i \in G$. ϵ is a small smoothness term that avoids devision by 0. The main region segmentation task, deep supervision task and the contour extraction task are set as Dice loss L_d^1, L_d^2 and L_d^3, respectively. The overall loss function L is denoted as following by summarizing the weighted losses:

$$L = \sum_{j=1}^{3} \lambda_j L_d^j + \beta \left\| W \right\|^2 \tag{2}$$

where λ_j denotes the task weights, β denotes the balance of weight decay term and W denotes the parameters of the whole network. To ensure that the region segmentation task dominates while other tasks take effects, λ is set as [1.0, 0.5, 0.5] respectively and β is set as 1e-5.

In the training phase, the gradient with respect to prediction $p_k \in P$ can be computed as following:

$$\frac{\partial L}{\partial p_k} = \sum_{j=1}^{3} -2\lambda_j \frac{\sum_{i=1}^{N} p_{i,j} g_{i,j} - g_{k,j} \sum_{i=1}^{N} (p_{i,j} + g_{i,j})}{[\sum_{i=1}^{N} (p_{i,j} + g_{i,j})]^2} \tag{3}$$

2.3. Multi-resolution model ensemble

In order to suppress the hard mimics around the cancer regions and preserve boundary details, we employed a multi-resolution model ensemble strategy. As is shown in Fig. 2, three networks of identical architecture are employed. HL-FCN1, HL-FCN2 and HL-FCN3 are fed with images of $4.0 \times 1.0 \times 1.0$ mm, $4.0 \times 1.5 \times 1.5$ mm and $4.0 \times 2.0 \times 2.0$ mm ZYX spacing, respectively. The cropped patches are firstly down-sampled to identical dimensions input into the networks, and subsequently predictions are up-sampled to original dimensions. While receptive fields in image domain keep identical, input patches of lower resolution enlarge the physical receptive field, i.e. the receptive field in world coordinate system, to feed more background context to the network, and patches of higher resolution preserves better boundary details. The ensemble stage is performed by averaging these predictions.

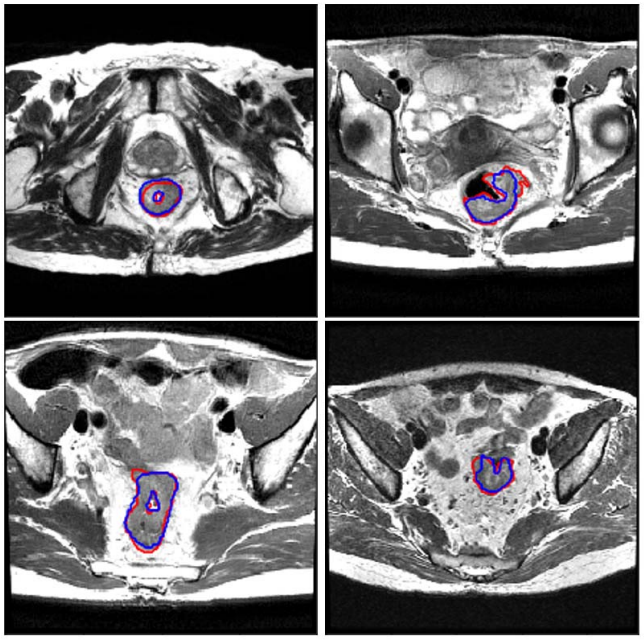

Fig. 3. Segmentation results of test cases. The red and blue contours denote the ground truth and our segmentation results, respectively.

3. EXPERIMENTAL RESULTS

3.1. Data set and preprocessing

The data set contains a total of 64 MR images of the pelvic cavity. Target areas were labeled voxel-wisely by experienced radiologists, and contour labels were automatically generated from the region labels using erosion and subtraction operations. The acquired MR images are T2 modality, and the ZYX spacing ranges from $3.6 \times 0.31 \times 0.31$ mm to $4.0 \times 1.0 \times 1.0$ mm. Four-fold cross-validation was conducted for reporting the performance.

The images are normalized and clipped to enhance the volume of interest. Firstly, body region was located using morphological operations. Then the normalization of intensity values was performed in the areas of body region. In addition, different transformations were conducted for data augmentation including translation, scale and intensity jittering.

Table 1. Comparison of colorectal cancer segmentation results using different methods.

Method	DSC	ASD[mm]
HL-FCN(Ensemble)	**0.721±0.139**	**3.83±4.95**
HL-FCN3	0.699±0.125	3.90±4.43
HL-FCN2	0.700±0.145	5.48±7.06
HL-FCN1	0.677±0.184	10.24±14.59
Dice-FCN(Ensemble)	0.699±0.137	4.18±5.89
Dice-FCN3	0.685±0.138	4.19±5.75
Dice-FCN2	0.673±0.153	5.70±7.31
Dice-FCN1	0.660±0.182	10.32±12.11
Ronneberger *et al.* [5]	0.617±0.192	4.26±4.35

3.2. Quantitative evaluation and comparison

We evaluate different methods using mean Dice Similarity Coefficient(DSC) and Average Surface Distance(ASD). ASD metric evaluates the shape similarity and is more sensitive to segmentation failures such as false positives and false negatives.

To investigate the contribution of each strategy adopted in our method, ablation studies are performed. FCNs of the same architecture with different loss formulations and ensemble strategies are reported in Table 1. By replacing the cross-entropy loss function with Dice loss and deep supervision in Dice-FCNs, the performance has been improved from the mean DSC from 0.617 to 0.660. By adding Dice-based contour-aware segmentation tasks to Dice-FCNs, the HL-FCNs gain further performance improvements. Finally, the multi-resolution ensemble strategy demonstrates its effectiveness by outperforming all other networks, achieving the best DSC score and smallest ASD. Four typical examples of our segmentation prediction are shown in Fig. 3, which validated the efficacy of our method by achieving good consistency with experienced radiologists.

We compared our method HL-FCN ensemble with state-of-the-art method in Table 1. 3D U-Net [5] was trained towards the cross-entropy loss function. Note that the compared network was adjusted to fit the scale and spacing set of the training data, i.e., $24 \times 96 \times 96$ input patches and $4.0 \times 1.0 \times 1.0$ spacing, and the conventional convolutional layers are replaced by ResBlocks. The evaluation results in Table 1 highlight that our proposed method outperforms the state-of-the-art method by a significant margin.

4. CONCLUSIONS

In this paper, we propose a volume-to-volume hybrid-loss guided FCN architecture for automatic colorectal cancer region segmentation from MR images. The incorporation of Dice-based hybrid loss function resolves class-imbalance and low-contrast appearance issues, and boosts the performance by a significant margin. We also further improve the performance by adopting multi-resolution model ensemble, to suppress false positives and preserve boundary details. In addition, this strategy is inherently general and can be used in other tasks encountering similar challenges.

5. REFERENCES

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] Dwarikanath Mahapatra, Peter J Schuffler, Jeroen AW Tielbeek, Jesica C Makanyanga, Jaap Stoker, Stuart A Taylor, Franciscus M Vos, and Joachim M Buhmann, "Automatic detection and segmentation of crohn's disease tissues from abdominal mri," *IEEE Trans. on Med. Imaging*, vol. 32, no. 12, pp. 2332–2347, 2013.

[3] Benjamin Irving, Amalia Cifor, Bartłomiej W Papież, Jamie Franklin, Ewan M Anderson, Michael Brady, and Julia A Schnabel, "Automated colorectal tumour segmentation in dce-mri using supervoxel neighbourhood contrast characteristics," in *MICCAI*. Springer, 2014, pp. 609–616.

[4] Hao Chen, Qi Dou, Xi Wang, Jing Qin, Jack C. Y. Cheng, and Pheng Ann Heng, "3d fully convolutional networks for intervertebral disc localization and segmentation," in *International Conference on Medical Imaging and Virtual Reality*, 2016, pp. 375–382.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *MICCAI*. Springer, 2016, pp. 424–432.

[6] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng-Ann Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d mr images.," in *AAAI*, 2017, pp. 66–72.

[7] H. Chen, Q. Dou, L. Yu, J. Qin, and P. A. Heng, "Voxresnet: Deep voxelwise residual networks for brain segmentation from 3d mr images," *Neuroimage*, 2017.

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 565–571.

[9] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, and P. A. Heng, "Dcan: Deep contour-aware networks for object instance segmentation from histology images.," *Medical Image Analysis*, vol. 36, pp. 135–146, 2017.

[10] Q. Dou, L. Yu, H. Chen, Y. Jin, X. Yang, J. Qin, and P. A. Heng, "3d deeply supervised network for automated segmentation of volumetric medical images," *Medical Image Analysis*, 2017.