Gland Instance Segmentation by Deep Multichannel Side Supervision

Yan Xu[1,2], Yang Li[1], Mingyuan Liu[1], Yipei Wang[1], Maode Lai[3], and Eric I-Chao Chang[2(✉)]

[1] State Key Laboratory of Software Development Environment and Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education and Research Institute of Beihang University in Shenzhen, Beihang University, Beijing 100191, China
[2] Microsoft Research, Beijing 100080, China
echang@microsoft.com
[3] Department of Pathology, School of Medicine, Zhejiang University, Hangzhou, China

Abstract. In this paper, we propose a new image instance segmentation method that segments individual glands (instances) in colon histology images. This is a task called instance segmentation that has recently become increasingly important. The problem is challenging since not only do the glands need to be segmented from the complex background, they are also required to be individually identified. Here we leverage the idea of image-to-image prediction in recent deep learning by building a framework that automatically exploits and fuses complex multichannel information, regional and boundary patterns, with side supervision (deep supervision on side responses) in gland histology images. Our proposed system, deep multichannel side supervision (DMCS), alleviates heavy feature design due to the use of convolutional neural networks guided by side supervision. Compared to methods reported in the 2015 MICCAI Gland Segmentation Challenge, we observe state-of-the-art results based on a number of evaluation metrics.

Keywords: Instance segmentation · Fully convolutional neural networks · Deep multichannel side supervision · Histology image

1 Introduction

Recent progress in deep learning technologies has led to explosive development in machine learning and computer vision for building systems that have shown substantial improvements in a wide range of applications such as image classification [7,10] and object detection [4]. The fully convolutional neural networks (FCN) [8] enable end-to-end training and testing for image labeling; holistically-nested edge detector (HED) [14] learns hierarchically embedded multi-scale edge fields to account for the low-, mid-, and high- level information for contours and object boundaries. FCN performs image-to-image training and testing, a factor that has become crucial in attaining a powerful modeling and computational capability of complex natural images and scenes.

S. Ourselin et al. (Eds.): MICCAI 2016, Part II, LNCS 9901, pp. 496–504, 2016.
DOI: 10.1007/978-3-319-46723-8_57

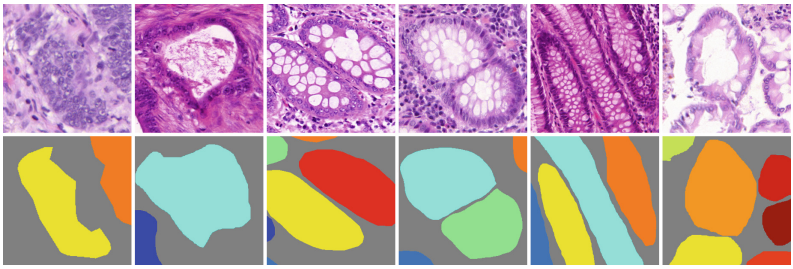

Fig. 1. Gland Haematoxylin and Eosin (H&E) stained slides and ground truth labels. Images in the first row exemplify different glandular structures. Characteristics such as heterogeneousness and anisochromasia can be observed in the image. The second row shows the ground truth. To achieve better visual effects, each color represents an individual glandular structure.

FCN family models [8,14] are well-suited for image labeling/segmentation in which each pixel is assigned a label from a pre-specified set. However, they can not be directly applied to the problem where individual objects need to be identified. This is a problem called instance segmentation. In image labeling, two different objects are assigned with the same label so long as they belong to the same class; in instance segmentation, objects belonging to the same class also need to be identified individually, in addition to obtaining their class labels.

Recent work developed in computer vision [2] shows interesting results for instance segmentation but a system like [2] is for segmenting individual objects in natural scenes. With the proposal of fully convolutional network (FCN) [8], the "end-to-end" learning strategy has strongly simplified the training and testing process and achieved state-of-the-art results in solving the segmentation problem back at the time. To refine the partitioning result of FCN, [6,15] integrate Conditional Random Fields (CRF) with FCN. However, they are not able to distinguish different objects leading to failure in instance segmentation problem. DCAN [1] and U-net [9] are two instance aware neural networks based on FCN.

The intrinsic properties of medical image pose plenty of challenges in instance segmentation [3]. First of all, objects being in heterogeneous shapes make it difficult to use mathematical shape models to achieve the segmentation. As Fig. 1 shows, when the cytoplasm is filled with mucinogen granule the nucleus is extruded into a flat shape whereas the nucleus appears as a round or oval body after secreting. Second, variability of intra- and extra- cellular matrix is often the culprit leading to anisochromasia. Therefore, the background portion of medical images contains more noise like intensity gradients, compared to natural images.

In this paper, we aim to developing a practical system for instance segmentation in gland histology images. We make use of multichannel learning [13], region and boundary cues using convolutional neural networks with side supervision, and solve the instance segmentation issue in the gland histology image. Our algorithm is evaluated on the dataset provided by MICCAI 2015 Gland Segmentation Challenge Contest [11,12] and achieves state-of-the-art performance.

2 Method

2.1 HED-Side Convolution (HED-SC)

The booming development of machine learning provides pathology slide image analysis with copious algorithms and tools. Although FCN has been shown to be excellent [8], due to the loss of boundary information during downsampling, FCN fails to distinguish instances in certain classes. To conquer this challenge, HED learns rich hierarchical representations under the guidance of deep supervision with each layer capable of carrying out an edge map at a certain scale. Thus the HED model is naturally multi-scale. Combining the side-outputs together, the weighted-fusion layer integrates the features obtained from different levels yielding superior results (for more details on HED, see [14]). Since our model performs the edge detection on the basis of pixelwise prediction, the transformation from the region feature to boundary feature is required. Hence, the original HED model is modified by adding two convolution layers in each side output path and the HED-SC model is born. We build a multichannel model (Fig. 2) that accomplishes the task of instance segmentation in the gland histology image.

2.2 Multichannel Learning

There are N images in the training set that can be divided into K categories. Note that K is the number of object categories. We denote our training set by $S = \{(X_n, Y_n, Z_n),\ n = 1, 2, ..., N\}$ where $X_n = \{x_j^{(n)},\ j = 1, 2, ..., |X_n|\}$ denotes the original input image, $Y_n = \{y_j^{(n)},\ j = 1, 2, ..., |Y_n|\}$, $y_j \epsilon \{0, 1, 2, ..., K\}$ and $Z_n = \{z_j^{(n)},\ j = 1, 2, ..., |Z_n|\}$, $z_j \epsilon \{0, 1\}$ denotes the corresponding ground truth label and binary edge map for image X_n respectively. X_n is simplified as X since all the training images are independent. Our goal is to predict the output set Y from the input image X. By multichannel, we emphasize that we exploit basic cues of segmenting images - region context and edge context - as two channels (Fig. 3).

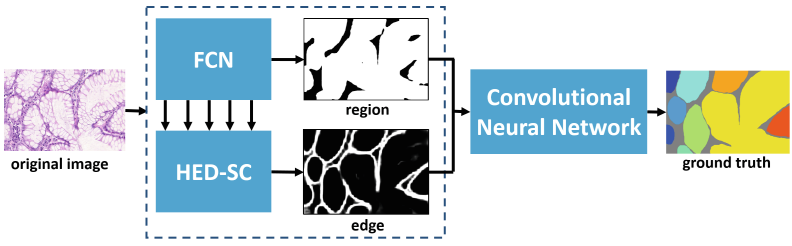

Fig. 2. This illustrates a brief structure of DMCS. FCN, the region channel, yields the prediction of regional probability maps. HED-SC, the edge channel, outputs the result of boundary detection. A convolution neural network concatenates features generated by different channels and produces segmented instances.

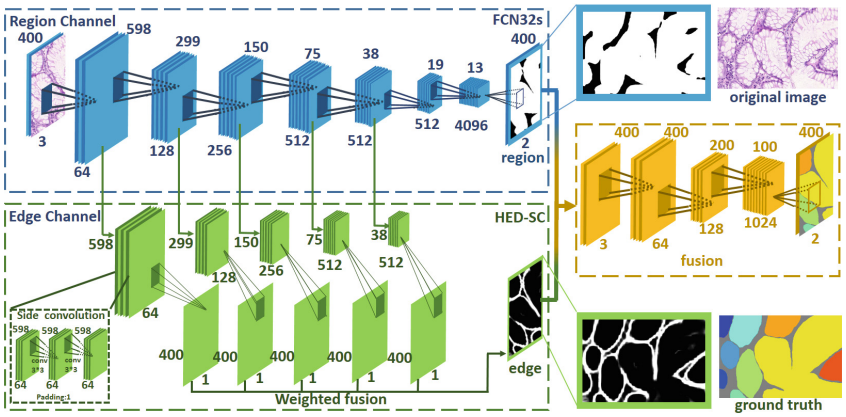

Fig. 3. Illustrates the deep multichannel side supervision model. The region channel engaged in producing a coarse pixel prediction of which the structure is identical to FCN32s [8]. At the first convolutional layer, padding of 100 pixels is involved as Long does [8]. The output of this channel achieved via the strategy of in-network up-sample layers and crop layers is the same size as the input images. Boundary information is obtained by the HED-SC channel of DMCS inspired by HED [14]. In this edge detection model side convolution is inserted before all the pooling layers in the FCN32s. Altogether, there are five side convolutions. Learnable weighting is assigned to five output of deep supervisions to produce the final result. The third part in DMCS aims to do instance segmentation based on information of region and boundary. It concatenates the output of the region channel and the HED-SC channel together. This fully convolutional neural network is utilized to process the segmented images.

Region feature channel. The region feature channel optimizes the pixel-wise prediction P_r. We fix the parameter w_e, w_f while learning the parameter w, w_r. The parameters in HED-SC and the parameters before the fully connection layer are represented as w_e and w_r respectively. Parameters in the fuse stage are denoted as w_f. Shared with both channels, the weights in FCN before w_r are represented as w. In this stage, our proposed model follows the architecture of FCN. Fully convolutional networks are trained pixel-to-pixel to achieve image semantic segmentation. Given an input image X, we first predict the pixel-to-pixel label Y^* where μ_k denotes the k^{th} class output of softmax function and $h(\cdot)$ calculates the activation of neural network:

$$P_r\left(y_j^* = k \mid X; w, w_r\right) = \mu_k\left(h\left(X, w, w_r\right)\right). \tag{1}$$

The loss function in this stage is the logarithmic loss function.

HED-SC channel. The HED-SC channel performs the edge detection on the pixel-wise prediction basis. First of all, the lower layer representation of most neural network lacks of semantic meaning due to the gradients vanishing/exploding problem during back-propagation. Deep supervised networks solve this exact problem by adding loss layers in lower structure of the network. In

our edge detection model, prior to each pooling layer, feature maps are executed with convolution operation with the kernel size of 3×3, yielding five heatmaps in this case. The prediction for each side-output is calculated as follows:

$$P_e^{(m)}\left(z_j^{*(m)} = 1 \mid X; w, w_e^{(m)}\right) = \sigma\left(h\left(X, w, w_e^{(m)}\right)\right). \tag{2}$$

$\sigma(\cdot)$ is the sigmoid function. Meanwhile, these five side-outputs are generated from feature maps with various sizes, in doing so the architecture of the network is naturally multi-scale. Weighted concatenating the five-scale side-outputs together (the weight $w_b^{(0)}$ is learnable), the low-, middle- and high-level information is integrated to generate the edge map:

$$P_e^{(0)}\left(z_j^{*(0)} = 1 \mid X; w, w_e\right) = \sigma\left(\sum_{m=1}^{M} w_e^{(0)(m)} \cdot h\left(X, w, w_e^{(m)}\right)\right). \tag{3}$$

The loss function of side-output and weighted fusion is cross entropy loss function thus the loss function of this chanel is the sum of these six losses. Merging side-outputs and weighted-fuse would optimize the edge detection result [14], but our priority is not edge detection thus we consider $P_e^{(0)}$ as the final edge prediction.

Training. At the training phase we combine the pixel prediction and edge prediction together and obtain the fine-grained pixelwise prediction Y_f^* as our final result:

$$P_f\left(y_{fj}^* = k \mid O_r, O_e^{(0)}; w_f\right) = \mu_k\left(h\left(O_r, O_e^{(0)}, w_f\right)\right), \tag{4}$$

where $O_r = h\left(X, w, w_r\right)$ and $O_e^{(0)} = \sum_{m=1}^{M} w_e^{(0)(m)} \cdot h\left(X, w, w_e^{(m)}\right)$. Firstly, it concatenates the output of first component, the pixel prediction, and the second component, the edge information, together. Then we apply a fully convolutional neural network to process the segmented images. This network contains four convolutional layers, two pooling layers, three full connected layers which are achieved by convolution and an up-sampling layer. We still choose the logarithmic loss function.

3 Experiment

Experiment data. The dataset is provided by MICCAI 2015 Gland Segmentation Challenge Contest [11,12] which consists of 85/80 labeled H&E stained colorectal cancer histological images in the training/testing sets (test A has 60 images and test B has 20 images).

Data augmentation. Followings are methods we deploy in augmentation: horizontal flipping, rotation operation (0, 90, 180, 270) and shifting operation.

Hyperparameters. CAFFE [5] is used in our experiments. Experiments are carried out on K40 GPU and the CUDA edition is 7.0. The weight decay is 0.002, the momentum is 0.9, and we choose 10 as the mini-batch size in order to use

GPU as efficient as possible. While training the region channel of the network, the learning rate is 10^{-3} and the parameters is initialized by pre-trained FCN32s model [8], while the HED-SC channel is trained under the learning rate of 10^{-9} and the Xavier initialization is performed. Fusion is learned under the learning rate of 10^{-3} and initialized by Xavier initialization. Finally, the whole framework is fine-tuned with the learning rate 10^{-3} and the weight of loss of edge is 10^{-6}.

Evaluation. Three criteria are engaged to evaluate results of instance segmentation. The summation of ranking numbers on two testing datasets determine the final ranking. The F1 score measures the accuracy of glandular instance segmentation. The true positive is defined as the segmented object which at least 50 % intersects with the ground truth. ObjectDice assesses the performance of segmentation. ObjectHausdorff evaluates the shape similarity between ground truth and segmented object based on object-level Hausdorff distance.

Result. Our framework performs well in the dataset provided by MICCAI 2015 Gland Segmentation Challenge and achieves state-of-the-art results (as listed in Table 1) among all participants [11]. We train FCN for 20 epoches in roughly 23 h, HED for 20 epoches in 22 h, the fusion phase for 10 epoches in 5 h and the finetune phase for 40 epoches in 50 h. The time of training and testing one image are 4-day and 1.5 s respectively. Compared to FCN our framework obtains better score which is a convincing evidence that our work is more effective in solving instance segmentation problem in histological images. Results of instance segmentation are illustrated in Fig. 4. Based on FCN, we add the region information to solve the instance segmentation task. Compared to FCN, most of the adjacent glandular structures have been separated apart which indicates that our framework accomplishes the instance segmentation goal. However, glands which are too small and have similar backgrounds (5th row in Fig. 4) are neither detected by FCN nor recognized in the fusion process. Images scattered with

Table 1. Our framework performs outstandingly in datasets provided by MICCAI 2015 Gland Segmentation Challenge Contest and achieves the state-of-the-art result. We rearrange the scores and ranks in this table. Our method outranks FCN and other participants [11] based on rank sum.

Method	F1 Score				ObjectDice				ObjectHausdorff				Rank Sum
	Part A		Part B		Part A		Part B		Part A		Part B		
	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	Score	Rank	
FCN	0.709	11	0.708	5	0.748	11	0.779	7	129.941	12	159.639	6	52
Ours	0.858	8	**0.771**	**1**	0.888	2	**0.815**	**1**	54.202	2	**129.930**	**1**	15
CUMed- [1] Vision2	**0.912**	**1**	0.716	4	**0.897**	**1**	0.781	6	**45.418**	**1**	160.347	8	21
ExB1	0.891	4	0.703	6	0.882	5	0.786	3	57.413	7	145.575	2	27
ExB3	0.896	2	0.719	3	0.886	3	0.765	8	57.350	6	159.873	7	29
Frerburg2 [9]	0.870	5	0.695	7	0.876	6	0.786	4	57.093	4	148.463	4	30
CUMed-Vision1 [1]	0.868	6	0.769	2	0.867	9	0.800	2	74.596	9	153.646	5	33

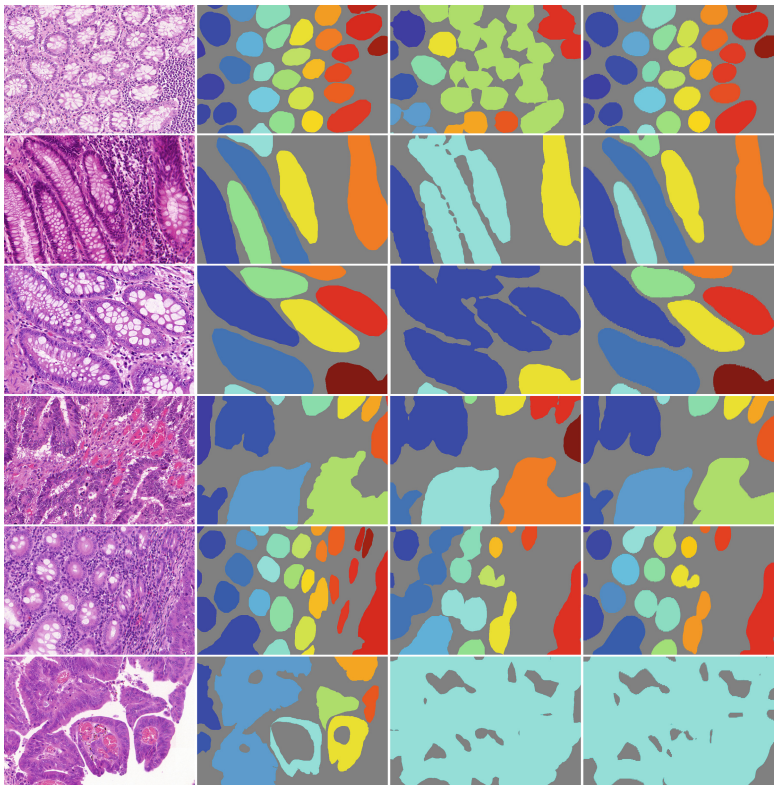

Fig. 4. From left to right: original image, ground truth, result using FCN, result using DMCS model. Compared to FCN, most of adjacent glandular structures are separated apart which indicates that our framework accomplishes the instance segmentation goal. However, few glands with small sizes or filled with red blood cells escape the detection of our model. The bad performance in the last row is due to the fact that in most samples, the white area is recognized as cytoplasm while in this sample, the white area is the background.

red blood cells caused by internal hemorrhage are excluded in training dataset, consequently instance segmentation result (6th row in Fig. 4) is not satisfactory.

Discussion. This framework exploits information from both region and edge channels, of which the region channel accomplishes the segmentation and positioning while the edge channel separates two adjacent gland instances. In test A, most images are the normal ones while test B contains a majority of cancerous images which are more complicated in shape and lager in size. Hence, a larger receptive field is required in order to detect cancerous glands. We use 5 pooling layers to enlarge the receptive field but in doing so, the network produces a much smaller heatmap (32 times subsampling of the original image) thus the performance concerning detecting small normal glands gets worse.

4 Conclusion

We propose a new algorithm called deep multichannel side supervision which achieves state-of-the-art results in MICCAI 2015 Gland Segmentation Challenge. The universal framework extracts features of both the edge and region and concatenate them together to generate the result of instance segmentation.

In future work, this algorithm can be utilized in instance segmentation of medical images.

Acknowledgement. This work is supported by Microsoft Research under the eHealth program, the Beijing National Science Foundation in China under Grant 4152033, Beijing Young Talent Project in China, the Fundamental Research Funds for the Central Universities of China under Grant SKLSDE-2015ZX-27 from the State Key Laboratory of Software Development Environment in Beihang University in China.

References

1. Chen, H., Qi, X., Yu, L., Heng, P.A.: Dcan: deep contour-aware networks for accurate gland segmentation. In: CVPR (2016)
2. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: CVPR (2016)
3. Dimopoulos, S., Mayer, C.E., Rudolf, F., Stelling, J.: Accurate cell segmentation in microscopy images using membrane patterns. Bioinformatics **30**(18), 2644–2651 (2014)
4. Girshick, R.: Fast r-cnn. In: ICCV (2015)
5. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: convolutional architecture for fast feature embedding. In: ICM (2014)
6. Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
8. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
9. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24574-4_28
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
11. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: the glas challenge contest. arXiv:1603.00275 (2016)
12. Sirinukunwattana, K., Snead, D.R., Rajpoot, N.M.: A stochastic polygons model for glandular structures in colon histology images. TMI **34**(11), 2366–2378 (2015)

13. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
14. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
15. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: CVPR (2015)