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
[2] 100191, Microsoft Research Asia, Beijing 100080, China,
echang@microsoft.com
[3] Zhejiang University, Hangzhou, Zhejiang 310058, 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 improvement 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.

⋆ Corresponding author.

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] and [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 with acceptant performance.

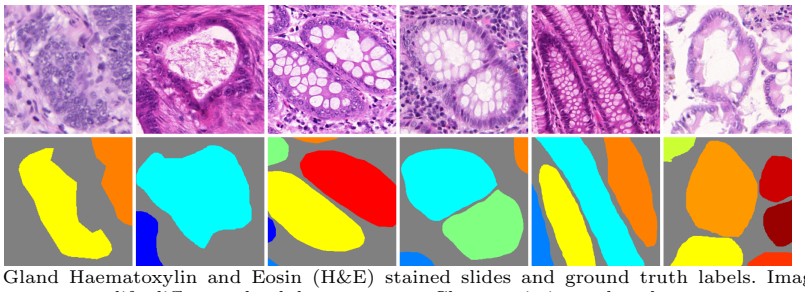

Fig. 1: Gland Haematoxylin and Eosin (H&E) stained slides and ground truth labels. Images in the first row exemplify different glandular structures. Characteristics such as heterogeneousness and anisochromasia can be observed in the image. The second row shows the ground truth. To achieve better visual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

The intrinsic properties of medical image pose plenty of challenges in instance segmentation [3]. First of all, the objects are in heterogeneous shapes, which make it difficult to use mathematical shape models to achieve the segmentation task. Take colorectal cancer histology image as an example (Fig.1). 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 of the medical image 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 engage 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 task of pathology image analysis is challenging yet crucial. 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 class. 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. In this paper, we build a multichannel model (Fig.2) that accomplishes the task of instance segmentation in the gland histology image.

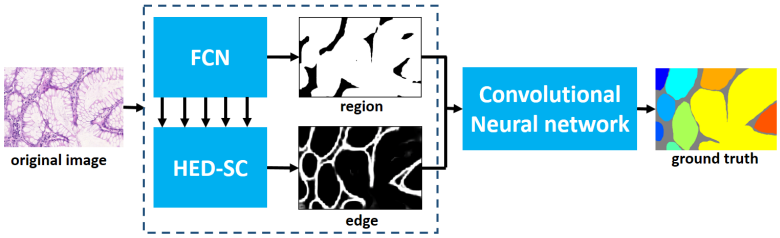

Fig. 2: Figure above illustrates a brief structure of DMCS. The black arrows represent the forward learning progress. FCN, the region channel, yields the prediction of regional probability maps. HED-SC, the edge channel, outputs the result of boundary detection. Convolution neural network is engaged to concatenate features generated by different channels and produce segmented instances.

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 plus. 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. For convenience, 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.

Region feature channel The region feature channel optimizes the pixelwise 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

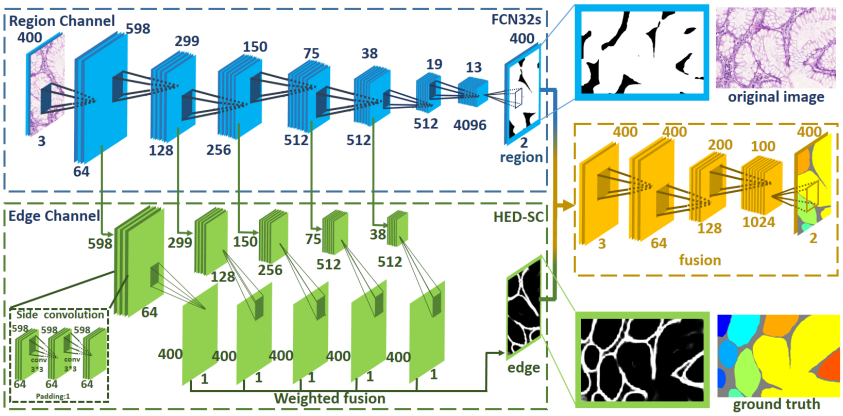

Fig. 3: Illustrates the deep multichannel side supervision model. The region channel engaged in producing a coarse pixel prediction of which the structure is identical to FCN32s [8]. At the first convolutional layer, padding of 100 pixels is involved as Long does [8]. The output of this channel achieved 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 [7]. 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.

denoted as w_f. Shared with both channels, the weights in FCN before w_r are symbolized 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 are

$$L_r\left(Y^*, X, w, w_r\right) = \sum_{j=1}^{|Y^*|} l_{log}\left(P_r\left(y_j^* = y_j \mid X; w, w_r\right)\right). \tag{2}$$

$l_{log}(\cdot)$ 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 lack of semantically meaning due to the gradients vanishing/exploding problem during back-propagation. Deep supervised networks solve the exact problem by adding loss layers in lower structure of 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:

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

$\sigma(\cdot)$ is the sigmoid function. The loss function for side-output is:

$$L_e^{(m)}\left(Z^*, X, w, w_e^{(m)}\right) = \sum_{j=1}^{|Z^*|} l_E\left(P_e^{(m)}\left(z_j^{*(m)} = 1 \mid X; w, w_e^{(m)}\right)\right), \quad (4)$$

$l_E(\cdot)$ is cross entropy loss 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:

$$P_e^{(0)}\left(z_j^{*(0)} = 1 \mid X; w, w_e\right) = \sigma\left(\sum_{m=1}^{M} w_e^{(0)(m)} \cdot h\left(X, w, w_e^{(m)}\right)\right), \quad (5)$$

and the loss function is

$$L_e^{(0)}\left(Z^*, X, w, w_e\right) = \sum_{j=1}^{|Z^*|} l_E\left(P_e^{(0)}\left(z_j^{*(0)} = 1 \mid X; w, w_e\right)\right), \quad (6)$$

Our loss function of this stage can be computed as

$$L_e\left(Z^*, X, w, w_e\right) = \sum_{m=1}^{M} L_e^{(m)}\left(Z^*, X, w, w_e^{(m)}\right) + L_e^{(0)}\left(Z^*, X, w, w_e\right), \quad (7)$$

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:

$$P_f\left(y_{fj}^* = k \mid O_r, O_e^{(0)}; w_f\right) = \mu_k\left(h\left(O_r, O_e^{(0)}, w_f\right)\right), \quad (8)$$

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:

$$L_f\left(Y_f^*, O_r, O_e^{(0)}, w_f\right) = \sum_{j=1}^{|Y_f^*|} l_{log}\left(P_f\left(y_{fj}^* = y_j \mid O_r, O_e^{(0)}; w_f\right)\right), \quad (9)$$

3 Experiment

Experiment data The dataset is provided by MICCAI 2015 Gland Segmentation Challenge Contest [11, 12] which consists of 165 labeled H&E stained colorectal cancer histological images. There are 85 images in the training set and 80 in the test sets (test A has 60 images and test B has 20 images).

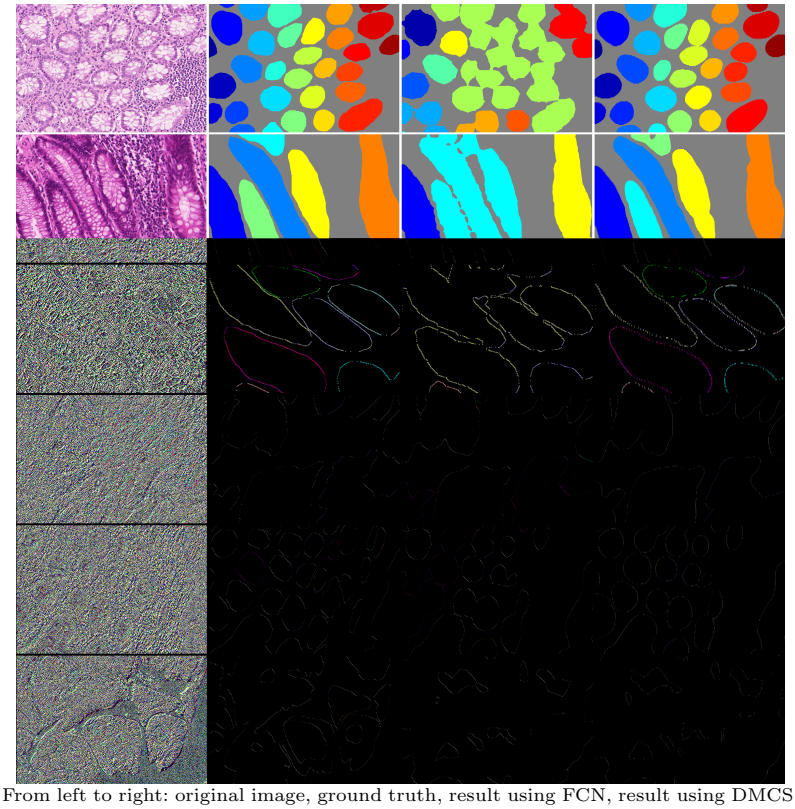

Fig. 4: From left to right: original image, ground truth, result using FCN, result using DMCS model. Compared to FCN, most of the adjacent glandular structures are separated apart which indicates that our framework accomplishes the instance segmentation goal. However, few glands with small shape or filled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

Data augmentation We first preprocess the data by performing per channel zero mean. To enhance performance and combat overfitting, copious training data are needed to learn the parameters. Given the circumstance of the absence of a large dataset, data augmentation is essential before training. The following lists five methods we deploy in augmentation. Horizontal flipping is used in our given dataset. The insensitivity of orientation in the gland slide enables the rotation operation (0,90,180,270) to training images. Meanwhile, shifting operation is applied to the available training images as well.

Hyperparameters We implement our learning network using a deep learning framework CAFFE [5]. Experiments are carried out on K40 GPU and the CUDA edition is 7.0. During the training phase, a back progression training strategy is involved. The parameters of the framework are as follows: weight decay is 0.002, momentum is 0.9, mini-batch size is 10. While training the region channel of the network, the learning rate is 10^{-3} and the parameters in the framework 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 the result of instance segmentation. The summation of six ranking numbers of three criteria on two testing datasets determine the final ranking of each team. 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.

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
CUMedVision2 [1]	**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
CUMedVision1 [1]	0.868	6	0.769	2	0.867	9	0.800	2	74.596	9	153.646	5	33

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.

Result Our framework performs well in the dataset provided by challenge of 2015 MICCAI and achieves state-of-the-art results (as listed in Table. 1) among all participants [11]. We train FCN 20 for epoches with approximately 23h, HED for 20 epoches with 22h and the fusion phase for 40 epoches with 50h. Compared to the result of 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.

The result of instance segmentation is illustrated in Fig.4. Our method is inspired by FCN and 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 (fifth row in Fig.4) are neither detected by FCN nor recognized in the fusion process. Images scattered with red blood cells caused by internal hemorrhage are excluded in training dataset, consequently instance segmentation result (sixth row in Fig.4) is not satisfactory.

Discussion This framework exploits information from both region and gland channels, of which the region channel accomplishes the segmentation and positioning while the edge channel separates two adjacent gland instances.

In test A, most of the pathology slide images are the normal ones while test B contains a majority of the images of cancerous tissue 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 medical images and multichannel learning can be used to improve instance segmentation.

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. We thank Zhuowen Tu for providing a great deal of help and support.

References

1. Chen, H., Qi, X., Yu, L., Heng, P.A.: Dcan: Deep contour-aware networks for accurate gland segmentation. arXiv preprint arXiv:1604.02677 (2016)
2. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. arXiv preprint arXiv:1512.04412 (2015)
3. Dimopoulos, S., Mayer, C.E., Rudolf, F., Stelling, J.: Accurate cell segmentation in microscopy images using membrane patterns. Bioinformatics pp. 2644–2651 (2014)
4. Girshick, R.: Fast r-cnn. In: ICCV. pp. 1440–1448 (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. arXiv preprint arXiv:1408.5093 (2014)
6. Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. 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. pp. 3431–3440 (2015)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015, pp. 234–241. Springer (2015)
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 preprint arXiv:1603.00275 (2016)
12. Sirinukunwattana, K., Snead, D.R., Rajpoot, N.M.: A stochastic polygons model for glandular structures in colon histology images. T-MI pp. 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. pp. 1–9 (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. pp. 1529–1537 (2015)