Panoptic Segmentation with an End-to-End Cell R-CNN for Pathology Image Analysis

Donghao Zhang[1(✉)], Yang Song[1], Dongnan Liu[1], Haozhe Jia[2], Siqi Liu[1], Yong Xia[2], Heng Huang[3], and Weidong Cai[1]

[1] School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia
dzha9516@uni.sydney.edu.au

[2] School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China

[3] Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, USA

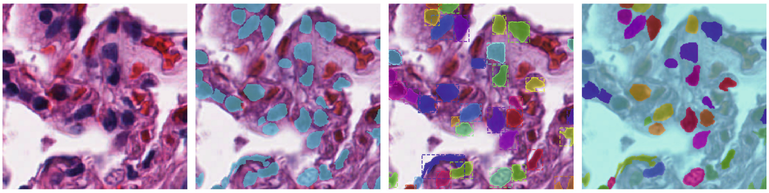
Abstract. The morphological clues of various cancer cells are essential for pathologists to determine the stages of cancers. In order to obtain the quantitative morphological information, we present an end-to-end network for panoptic segmentation of pathology images. Recently, many methods have been proposed, focusing on the semantic-level or instance-level cell segmentation. Unlike existing cell segmentation methods, the proposed network unifies detecting, localizing objects and assigning pixel-level class information to regions with large overlaps such as the background. This unifier is obtained by optimizing the novel semantic loss, the bounding box loss of Region Proposal Network (RPN), the classifier loss of RPN, the background-foreground classifier loss of segmentation Head instead of class-specific loss, the bounding box loss of proposed cell object, and the mask loss of cell object. The results demonstrate that the proposed method not only outperforms state-of-the-art approaches to the 2017 MICCAI Digital Pathology Challenge dataset, but also proposes an effective and end-to-end solution for the panoptic segmentation challenge.

1 Introduction

Cancer diagnosis by pathologists mainly relies on the visual inspection of tissue sample images captured by microscopy. The morphological features of cells such as the shape and nuclei size are significant to the diagnosis of the cancer stages (benign and malignant). It is impractical and labour-intensive for pathologists to produce manual morphological annotations for the whole slide image.

The general semantic scene understanding can be categorized into semantic segmentation, object detection, instance segmentation, and panoptic segmentation, as shown in Fig. 1. The instance segmentation assigns pixel-level segmentation masks to each individual object. The semantic segmentation obtains pixel-level whole image classification without differentiating different objects belong

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11071, pp. 237–244, 2018.
https://doi.org/10.1007/978-3-030-00934-2_27



(a) Image (b) Semantic Segmentation (c) Instance Segmentation (d) Panoptic Segmentation

Fig. 1. An example of illustrating the difference between semantic segmentation, instance segmentation and panoptic segmentation.

to the same class, but panoptic segmentation produces its semantic label (class) and its instance id. Recently in general computer vision, semantic segmentation has drawn a lot of attention, and many methods [1,10,16] based on convolutional neural network architectures are proposed. A U-shaped neural network [12] further includes more feature channels in the upsampling layer, which is proven to generate reasonable segmentation from limited training data such as neuronal structure segmentation of electron microscopy images. LinkNet [1] accelerates the processing time and reduces the network parameters by utilizing addition of feature channels instead of concatenation. Besides semantic segmentation, breakthroughs of object detection and instance segmentation happen due to the region proposal strategy by Faster R-CNN [11] and mask head segmentation by Mask R-CNN [5,6]. The recently introduced panoptic segmentation [9] defines its meaning by unifying semantic segmentation and instance segmentation.

In addition to recent developments of semantic scene understanding in general computer vision, there have been attempts [2,13–15] particularly targeting cell segmentation. The major difference between bio-medical related segmentation and general computer vision is the limited training data due to the difficulty of manual labelling requiring the prior knowledge of the specialists. Some proposed to use unannoted images with adversarial network [15] and others attempt to include contour-aware loss [2] and suggestive annotation [13] to solve these biomedical segmentation problems. Cell segmentation requires a unique id for each cell so only semantic segmentation is not enough to produce individual object segmentation masks. Although instance segmentation can produce unique ids for each cell object, it only relies on image features of intracellular materials. Moreover, the panoptic segmentation not only produces instance cell masks but also fully utilizes the potential information relation between intercellular and intracellular materials.

Inspired by the related work mentioned above, we propose an end-to-end panoptic segmentation method, named Cell R-CNN, to perform morphological analysis of pathology images. Our contribution is three-fold. First, we propose a unified solution of combining semantic segmentation and instance segmentation by introducing the novel semantic segmentation branch. Second, the proposed branch is capable of detecting regions having large overlaps with objects.

Third, we propose to use upsampling layers to replace deconvolution layers in GCN [10] to reduce the required parameters. Evaluated on the 2017 MICCAI Digital Pathology Challenge dataset, results indicate that the proposed Cell R-CNN outperforms state-of-the-art methods in terms of segmentation metrics including F1-score, Dice, and Hausdorff distance.

2 Methods

The main idea of the proposed framework is to unify semantic segmentation and instance segmentation. The convolutional features like ResNet [4] are used by the RPN [11] in object detection to reduce the computation time of the proposal. Similarly, some recent advances in semantic segmentation also demonstrate that the accuracy of semantic segmentation is improved by applying ResNet as its encoder. Is that possible to share these convolutional features between semantic segmentation and instance segmentation? In this paper, we simply and intuitively choose the global convolutional network (GCN) [10] as our base of semantic segmentation branch. The proposed Cell R-CNN firstly generates the convolutional features using the backbone ResNet, which are shared by the novel and intuitive semantic segmentation branch and feature map branch. This sharing operation improves the potential feature representation ability of backbone network (ResNet). The outputs of feature map branch are then fed into RPN to generate proposals. The region-of-interest proposals are the inputs of the instance segmentation branches [5]. Finally, the multi-task losses are optimized together to obtain the panoptic segmentation result.

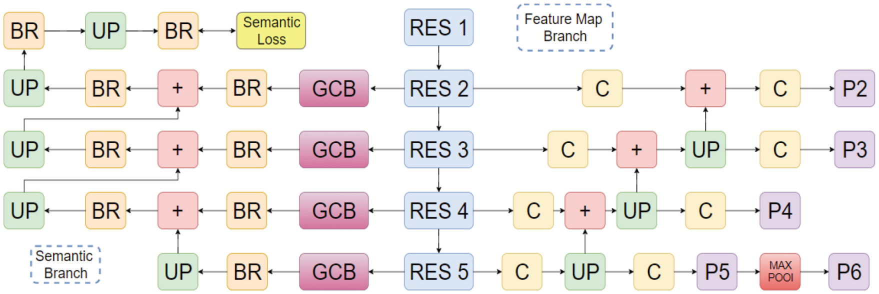


Fig. 2. The network architecture of semantic segmentation branch and feature map branch. C, BR, UP, GCB and RES X represent convolutional layer, boundary refine block, upsampling layer, global convolution block and specific ResNet layer respectively.

2.1 Semantic Segmentation Branch and Feature Map Branch

The overview of the semantic segmentation branch and the feature map branch is shown in Fig. 2. The semantic segmentation branch generates segmentation by assigning pixel-level object class information and the feature map branch prepares the features shared by RPN and instance branches (Sect. 2.2). GCN is chosen for the semantic segmentation branch because it is highly efficient and has global receptive field. ResNet101 is used as the backbone of GCN. The GCB employs convolutional kernels with the size $1 \times k + k \times 1$ and $k \times 1 + 1 \times k$ to ensure that there is enough valid receptive field. Instead of directly using the convolutional kernels with $k \times k$, the GCB is designed to greatly reduce the required training parameters for the network. BR consists of a $conv{+}relu{+}conv$ block and a residual design: $x = F(x) + x$. BR aims to replace non-trainable conditional random field or other post-processing techniques. In simple words, everything is learnable. Besides the semantic segmentation branch, the features generated by ResNet are also shared by the feature map branch. This design not only maximizes the potential feature representation ability of ResNet but also greatly reduces the required parameters. The feature map branch targets at sharing features $(P2, P3, P4, P5, P6)$ between RPN and the Instance Branch (Sect. 2.2). The first and second the convolutional kernel sizes of feature map branch after ResNet X are 1×1 and 3×3, respectively.

2.2 Region Proposal Network and Instance Branch

The inputs of RPN are the feature maps (P2, P3, P4, P5, P6) and its outputs are bounding box proposals with a score showing the possibility of being an object. The anchors (rectangular bounding boxes) are initially generated by sliding window strategy. At each sampling point, anchors with different ratios and sizes are generated. Each anchor is assigned a score and box delta $(t_x^*, t_y^*, log(w/w_a), log(h/h_a))$ by the RPN shown in Fig. 3. The bounding box regression [3] is defined with the following equations:

$$\begin{aligned} &t_x = (x - x_a)/w_a,\ t_y = (y - y_a)/h_a,\ t_w = log(w/w_a),\ t_h = log(h/h_a \\ &t_x^* = (x^* - x_a)/w_a,\ t_y^* = (y^* - y_a)/h_a,\ t_w^* = log(w^*/w_a),\ t_h^* = log(h^*/h_a) \end{aligned} \tag{1}$$

where x, y, w and h represent the coordinates of the center sampling points, the width and height of the predicted bounding box (ROIs). Similarly, x_a, y_a, w_a and h_a denote corresponding variables of the anchor box. The other variables are for the ground truth. The outputs of RPN are further refined by the proposal layer. The proposal layer sorts anchors by scores decreasingly. The box delta refinement is then applied and non-max suppression removes candidates of refined bounding boxes with strong overlaps with each other. For optimization of boxes and scores loss, only the positive anchors contribute to the loss calculation. In order to successfully train RPN, the ratio of positive anchors to negative anchors is maintained within a certain range. The kernel sizes of first and last RPN convolution layers are 3×3 and 1×1, respectively.

The instance branches consist of instance location and discriminator branch and instance mask branch. The instance location and discriminator branch further refine bounding box location and evaluate the possibility of each object category (foreground or background), which is a two-stage detector. The instance

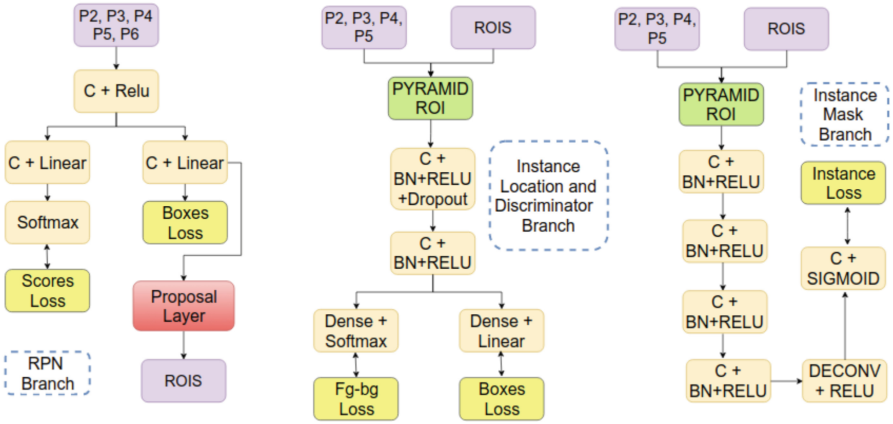


Fig. 3. Region proposal network and instance branch. ROIS, PYRAMID ROI, and DECONV indicate regions of interest, pyramid ROI aligning layer and deconvolutional layer, respectively.

branches use refined ROIs generated by the proposal layer as part of its input, and the other inputs are the feature maps (P2, P3, P4, P5). The pyramid ROI aligning layer selects the corresponding level of feature map based on the size of ROI. Here, instead of class specific loss, we use foreground-background loss. This particular design attempts to segment cells from different classes using the same network inspired by the segment everything work [6]. The instance mask branch is to generate an instance object segmentation when an ROI is given. For kernel sizes of convolutional layer for instance mask branch, all of them are 3×3 except the last convolutional layer is 3×3. The kernel size of deconvolutional layer is 2×2 with stride 2. For the computation of instance loss, the ground truth of object mask is resized into 28×28. This design is to preserve the spatial relation information. Due to pixel-to-pixel correspondence of pyramid ROI aligning layer, the prediction result of instance mask branch can be accurate.

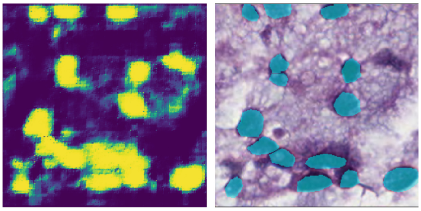
The final loss is defined with the following equation:

$$\begin{aligned} L_{total} = L_{semantic} + L_{score} + L_{boxes}(RPN) \\ + L_{(fg-bg)} + L_{boxes}(Instance) + L_{instance} \end{aligned} \tag{2}$$

where L_{total}, $L_{semantic}$, L_{score}, $L_{boxes}(RPN)$, $L_{(fg-bg)}$, $L_{boxes}(Instance)$, and $L_{instance}$ represent the total loss of Cell R-CNN, semantic segmentation loss, classifier loss of RPN, bounding box regression loss of RPN, foreground-background loss of instance branch, bounding box regression loss of instance branch, and mask segmentation loss of each object respectively as shown in Figs. 2 and 3; $L_{semantic}$ and $L_{instance}$ are categorical cross-entropy loss and binary cross-entropy loss respectively; $L_{boxes}(RPN)$ and $L_{boxes}(Instance)$ are the smoothed $L1$ regression losses, $L1(t - t^{*})$. L_{score} and $L_{(fg-bg)}$ are simple log losses between different classes.

3 Experiment

Experimental evaluation was performed on the 2017 MICCAI Digital Pathology Challenge dataset. The dataset is composed of 32 training images and 32 testing images. Both the training and testing images were sampled from the whole slide image of patients with glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSCC), lower grade glioma (LGG) tumors or non small cell lung cancer (NSCLC). The number of each category is 8 in both the training and testing datasets. The image size is either 500 × 500 or 600 × 600.



(a) Semantic Branch (b) Ground Truth

Fig. 4. An example of semantic segmentation branch result.

The data augmentation technique was applied to both the proposed method and comparison methods. The proposed method is implemented using keras and tensorflow. The convolutional kernel size for GCN block of semantic segmentation branch is constant as 5. The upsampling size along x, y dimension of semantic segmentation branch is 2. Since the maximum number of cells is less than 100, the number of training regions of interest is set to 400 to increase potential candidates. The pixel-size of the anchors are {8, 16, 32, 64, 128}. The stride of the anchors is 2. The optimizer is stochastic gradient descent (SGD) whose initial learning rate, momentum and weight decay are $2e^{-3}$, 0.9 and $1e^{-4}$, respectively. The non-maximum suppression threshold is set to 0.3 to ensure the successful detection of boundary-touching cells. The minimum detection confidence is set to 0.5.

An example of semantic branch result is shown in Fig. 4. The semantic branch is able to learn the region with large overlaps with other objects. In this particular example, the non-cell background region overlaps with individual cell objects but successfully distinguished by the semantic segmentation branch. One challenging condition is that only limited training data was provided while various cancer categories lead to different imaging conditions shown in Fig. 5. Overall, the proposed method is capable of producing accurate and reliable panoptic segmentation of different cancer images. According to the segmentation results, there are still incomplete cell segmentation of prediction result such as top left of prediction result on NSCLC image in Fig. 5. For quantitative evaluation, F1 score, object-level Dice, and Hausdorff distance are computed. The detailed definitions of F1 score, object-level Dice, and Hausdorff distance are referred to [2].

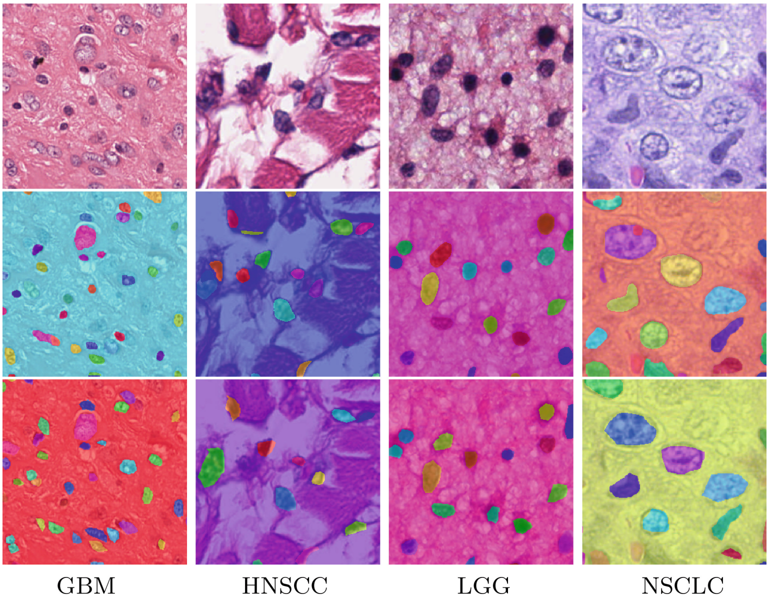


Fig. 5. Panoptic-level cell segmentation results of various cancer categories (from top to bottom): testing image, prediction, ground truth.

Table 1. The quantitative cell segmentation results.

Network	F1-score	Dice	Hausdorff
UNet [12]	0.4059 ± 0.2311	0.4942 ± 0.1872	54.1130 ± 73.4936
Pix2Pix [7]	0.6208 ± 0.1126	0.6351 ± 0.0706	19.1441 ± 6.0933
LinkNet [1]	0.4117 ± 0.1852	0.5611 ± 0.0899	19.7294 ± 9.0798
FnsNet [8]	0.7413 ± 0.0668	0.6165 ± 0.0839	25.9102 ± 9.5834
Ours w/o Semantic branch	0.8004 ± 0.0722	0.7070 ± 0.0598	12.6723 ± 3.4591
Ours	0.8216 ± 0.0625	0.7088 ± 0.0564	11.3141 ± 2.6917

It can be seen from Table 1 that the proposed method ranked first in terms of F1-score with 0.8216, Dice with 0.7088, and minimum Hausdorff distance with 11.3141. Based on the average and standard deviation of the evaluation metrics, the proposed framework is most robust and stable among all compared methods.

4 Conclusions

In this paper, we propose a novel and end-end Cell R-CNN framework to generate panoptic segmentation. The proposed method unifies individual semantic and instance segmentation tasks with a novel semantic segmentation branch. The semantic segmentation branch is capable of learning features from regions with large overlaps with other objects. Evaluated on the 2017 MICCAI Digital Pathology Challenge dataset, the proposed method outperforms compared methods in terms of F1 score, cell-object Dice, and Hausdorff distance.

References

1. Chaurasia, A., Culurciello, E.: Linknet: exploiting encoder representations for efficient semantic segmentation. arXiv preprint arXiv:1707.03718 (2017)
2. Chen, H., Qi, X., Yu, L., Heng, P.: DCAN: deep contour-aware networks for accurate gland segmentation. In: CVPR, pp. 2487–2496 (2016)
3. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 (2014)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV, pp. 2980–2988 (2017)
6. Hu, R., Dollár, P., He, K., Darrell, T., Girshick, R.: Learning to segment every thing. In: CVPR (2018)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR, pp. 5967–5976 (2017)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
9. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. arXiv preprint arXiv:1801.00868 (2018)
10. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters-improve semantic segmentation by global convolutional network. In: CVPR, pp. 1743–1751 (2017)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: NIPS, pp. 91–99 (2015)
12. 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, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
13. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: a deep active learning framework for biomedical image segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 399–407. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_46
14. Zhang, D., Song, Y., Liu, S., Feng, D., Wang, Y., Cai, W.: Nuclei instance segmentation with dual contour-enhanced adversarial network. In: ISBI (2018)
15. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 408–416. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_47
16. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR, pp. 6230–6239 (2017)