Nuclei Segmentation via a Deep Panoptic Model with Semantic Feature Fusion

Dongnan Liu[1*], **Donghao Zhang**[1*], **Yang Song**[2], **Chaoyi Zhang**[1], **Fan Zhang**[3], **Lauren O'Donnell**[3] and **Weidong Cai**[1]
[1]School of Computer Science, University of Sydney, Australia
[2]School of Computer Science and Engineering, University of New South Wales, Australia
[3]Brigham and Women's Hospital, Harvard Medical School, USA
{dliu5812, dzha9516}@uni.sydney.edu.au, yang.song1@unsw.edu.au, czha5168@uni.sydney.edu.au,
{fzhang, odonnell}@bwh.harvard.edu, tom.cai@sydney.edu.au

Abstract

Automated detection and segmentation of individual nuclei in histopathology images is important for cancer diagnosis and prognosis. Due to the high variability of nuclei appearances and numerous overlapping objects, this task still remains challenging. Deep learning based semantic and instance segmentation models have been proposed to address the challenges, but these methods tend to concentrate on either the global or local features and hence still suffer from information loss. In this work, we propose a panoptic segmentation model which incorporates an auxiliary semantic segmentation branch with the instance branch to integrate global and local features. Furthermore, we design a feature map fusion mechanism in the instance branch and a new mask generator to prevent information loss. Experimental results on three different histopathology datasets demonstrate that our method outperforms the state-of-the-art nuclei segmentation methods and popular semantic and instance segmentation models by a large margin.

1 Introduction

Cell morphology in histopathology images provides critical information for cancer diagnosis and prognosis. The first step in cell morphology analysis is the segmentation of individual cell nuclei, which is typically performed manually in current clinical practice. However, manual examination of histopathology images is labor-intensive and time-consuming due to the large image sizes and complex cellular structures. Therefore, investigating computerized methods is necessary to reduce the workload for pathologists, and make the analysis efficient [Veta *et al.*, 2014].

There are still some major challenges in nuclei segmentation tasks, as illustrated in Figure 1. First, there is a high level of heterogeneity in appearance between different types of organs or cells. Consequently, methods designed based on prior knowledge about geometric features cannot be applied directly to different images. Second, other structures such as cytoplasm and stroma can have similar features to the nuclei, making it hard to differentiate cell nuclei from the background. Third, nuclei are often clustered with many overlapping instances [Chen *et al.*, 2017]. In order to find the exact location and boundary for every single nucleus, further processing is often required to separate the clustered or overlapping nuclei.

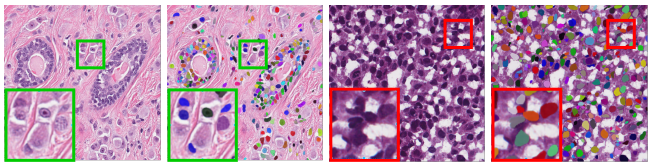

Figure 1: Example histopathology images for nuclei segmentation from different organs: breast (a) and corresponding annotation (b); bladder (c) and corresponding annotation (d); green boxes: cytoplasm which are similar to nuclei; red boxes: touching nuclei.

Convolutional neural networks (CNN) are powerful for tackling image recognition tasks [He *et al.*, 2016; Peng *et al.*, 2017] by learning the features automatically. Currently, most CNN related works for nuclei segmentation are based on the semantic segmentation model to separate foreground and background, and involve post-processing for refinement. Recently, instance segmentation models have been proposed for predicting the mask and region of interest (ROI) [He *et al.*, 2017; Liu *et al.*, 2018b]. When utilizing them on nuclei segmentation tasks, each ROI represents a single nucleus.

With semantic segmentation models [Peng *et al.*, 2017; Ronneberger *et al.*, 2015], all pixels are categorized into different classes, which are employed for studying the uncountable "stuff" of the image. By contrast, instance segmentation is able to assign unique labels to each object that belongs to the same class, and is therefore employed to study the countable "things". The study of both stuff and things is necessary for image recognition because the former obtains the features of the background while the latter is able to learn the features of different objects in the foreground. In order to reconcile the foreground and background learning, panoptic segmentation [Kirillov *et al.*, 2018] has been proposed to incorporate semantic segmentation with instance segmentation, with separate training of the instance and semantic branches.

Different from existing methods, in our model, we propose

[*]Authors contributed equally.

an end-to-end panoptic segmentation framework incorporating an auxiliary semantic segmentation branch with an instance branch which contains a dual-modal mask generator and feature fusion mechanism. First, we consider that semantic segmentation methods are only able to process global features, which limits their capacity to separate the individual nuclei. On the other hand, instance segmentation models focus more on the local ROI-level information and lack a global interpretation of the image. To address these limitations and also utilize the advantages of both the semantic and instance models, we propose to design a dual-branch panoptic segmentation model by integrating the semantic segmentation branch with the instance segmentation branch. The dual-branch model is trained end-to-end and is able to process the global and local features from the image at the same time. Furthermore, we introduce a new semantic feature map from the instance branch to further encourage the instance branch to encode the global features along with local ones during training. In order to prevent information loss when predicting the mask on each ROI, a dual-modal mask generator is designed with a fully connected (FC) mode and a spatial upscaling mode. Our main contributions can be summarized as follows:

- A novel dual-branch panoptic model is proposed with an instance segmentation branch and an auxiliary semantic segmentation branch. The semantic branch is specially designed for enlarging the receptive field of the input object and is jointly trained with the instance branch, in an end-to-end fashion.
- A feature fusion mechanism in the instance branch is designed to integrate global with local features.
- A dual-modal mask generator is designed for higher mask segmentation accuracy by minimizing the information loss.
- Our method was proven to be effective with significant improvements over the other state-of-the-art methods on three public datasets: TCGA-kumar [Kumar *et al.*, 2017], TNBC [Naylor *et al.*, 2018], and MICCAI 2017 Digital Pathology Challenge dataset [Vu *et al.*, 2018].

2 Related Work

2.1 Nuclei Segmentation

Studies into nuclei segmentation in histological images have been ongoing for many years. With progress in pattern recognition techniques, methods based on machine learning have shown encouraging results. These methods typically start with handcrafted feature extraction, such as textural features [Zhang *et al.*, 2014], Laplacian and Gaussian features [Kong *et al.*, 2013], and geometric features about contours [Wienert *et al.*, 2012]. Then, classification (e.g., Bayesian) or clustering (e.g., K-means) techniques are employed for nuclei segmentation and detection tasks [Naik *et al.*, 2008; Chankong *et al.*, 2014].

With the advance of CNN, nuclei segmentation has been modeled as a pixel- or patch-level classification problem. In [Raza *et al.*, 2019], a multi-resolution CNN is proposed

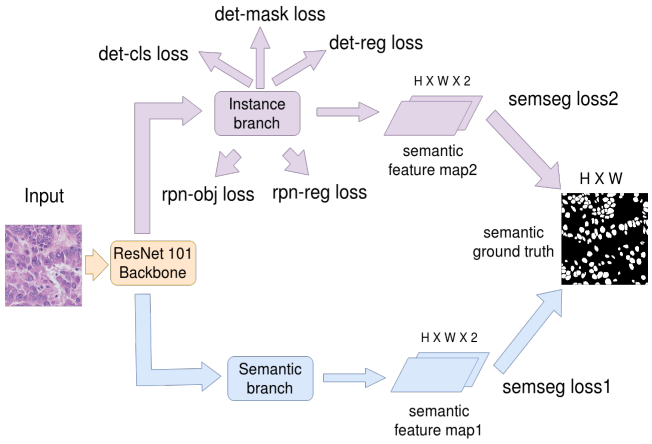

Figure 2: Overview of our proposed framework. Refer to Section 3.3 for the detailed loss definitions.

to process the nuclei with different receptive fields. Beyond the standard CNN model, improvements have been made to incorporate the contour information into the CNN architecture to facilitate segmentation between individual nuclei. For example, in [Kumar *et al.*, 2017], nuclei boundaries are considered as the third class for the CNN segmentation model. [Oda *et al.*, 2018] proposed a boundary enhanced module and loss function based on the traditional U-Net [Ronneberger *et al.*, 2015], and it facilitates the model's learning of more details about the nuclei that are hard to segment. Furthermore, Cell R-CNN [Zhang *et al.*, 2018a] achieves competitive performance simply by incorporating a semantic segmentation model with an instance model. In addition, regression-based models have also been proposed such as one uses the distance map as the ground truth labels [Naylor *et al.*, 2018], and another which applies general adversarial architecture [Zhang *et al.*, 2018b], which achieved competitive performance on nuclei segmentation tasks as well.

2.2 CNN Based Image Segmentation

In the field of semantic segmentation, skip connections between encoders and decoders are effective and prevalent [Ronneberger *et al.*, 2015; Zhang *et al.*, 2018c; Wang *et al.*, 2019]. In [Zhang *et al.*, 2018c] and [Liu *et al.*, 2018a], dense connections between the decoders in different resolutions are proposed to eliminate the information loss. In addition to semantic segmentation architectures, instance segmentation models such as [He *et al.*, 2017] and [Liu *et al.*, 2018b] are able to generate the masks of the image with a detection based architecture [Ren *et al.*, 2015]. Beyond semantic and instance segmentation, panoptic segmentation has been proposed to fuse the feature from things and stuff [Kirillov *et al.*, 2018]. In [Zhang *et al.*, 2018a] and [de Geus *et al.*, 2018], both the instance and semantic segmentation branches are trained together by sharing the same set of parameters in the backbone module. Then, the losses of the two branches are summed together for back propagation to optimize the parameters of the whole framework. In [Xiong *et al.*, 2019], a novel panoptic segmentation head is proposed to fuse the feature about

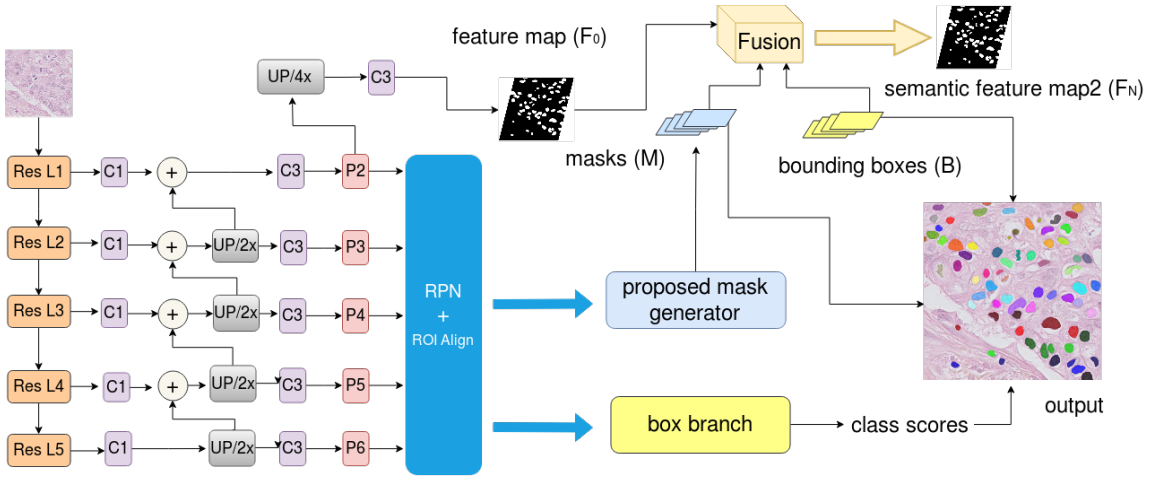

Figure 3: Illustration of the instance branch. *C1* and *C3* represent convolutional layers with a kernel size of 1 and 3, respectively. For *C3*, both the stride and padding size are 1. *UP/2x* and *UP/4x* represent a nearest unsampling layer with a scale factor of 2 and 4, respectively. We omit the *ReLU* layer after each convolutional block for brevity.

things and stuff, from instance and semantic branch, respectively. More recently, Panoptic Feature Pyramid Network was proposed in [Kirillov *et al.*, 2019] to generate a semantic output in the instance segmentation branch. The model achieves new state-of-the-art performance in several image segmentation tasks with higher memory efficiency.

3 Methods

The overall architecture of our work is illustrated in Figure 2. During training, both the semantic and instance branches are trained together with the same ResNet101 backbone module. During inference, only the instance branch is used to predict the class scores, bounding box coordinates, and masks. In this section, we present the designs of our network model.

3.1 Instance Branch

Figure 3 is a detailed illustration of our proposed instance branch for nuclei segmentation. In general, this branch is inspired by the structure of Mask R-CNN [He *et al.*, 2017], but with major differences in our design of feature fusion and the mask generator. Specifically, after the ResNet101 backbone, a Feature Pyramid Network (FPN) is applied to generate feature maps in five stages (*P2*, *P3*, *P4*, *P5*, *P6*). Next, with the multiple feature maps P2 to P6, the Region Proposal Network (RPN) and ROI Align modules are applied to obtain the ROIs for cell nuclei. Then, each ROI passes through a box branch for class score and bounding box coordinate prediction and our dual-modal mask generator for generating the nucleus masks. In addition, the feature map at the highest spatial resolution of FPN (*P2*) passes through an upsampling layer with a spatial factor of 4 and a convolutional layer with a channel number of 2. Then, this feature map with the same size as the semantic segmentation ground truth is fused with the nuclei masks according to their locations with feature fusion mechanism. The motivation is that in instance segmentation models, only local features such as intracellular detail and location for each single nucleus can be processed, which eliminates the global information processing. In this way, the global information is fused with the local information to enable more accurate localization of the cell nuclei.

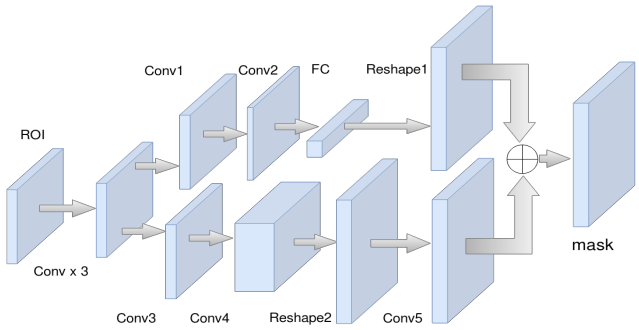

Figure 4: Illustration of the proposed mask generator. *Conv1*, *Conv2*, *Conv3*, *Conv4*, and *Conv5* represent five different convolutional layers, and *Conv* $\times$ 3 means three consecutive convolutional layers. *FC* is a fully connected layer, *Reshape1* contains reshaping and channel duplication, and *Reshape2* is a pixel shuffle. The final fusion is a pixel-wise summation. We omit the *ReLU* layer after each convolutional block for brevity.

Mask Generator

Figure 4 illustrates our proposed mask generator, which contains two parts. In the original Mask R-CNN, only four consecutive convolutional layers and a deconvolutional layer are employed to segment the mask. However, the size of the input ROI of the mask generator is 14×14, which is small, and directly applying a deconvolutional layer to such a small region is prone to cause information loss. On the other hand, an FC layer is able to learn the global information for the entire ROI, which is helpful to separate different instances with the same category label. Therefore, we incorporate the ideas of [Liu *et al.*, 2018b] and [Wang *et al.*, 2018], and design a mask generator which fuses the feature maps of dual modalities from the FC branch and spatial upscaling branch as illustrated in Fig-

Stage	Hyperparamaters	Output size
Input		$256 \times 14 \times 14$
Conv $\times$ 3	$k = (3,3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv1	$k = (3,3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv2	$k = (3,3), s = 1, p = 1$	$128 \times 14 \times 14$
FC		$1 \times 1 \times 784$
Reshape1		$2 \times 28 \times 28$
Conv3	$k = (3,3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv4	$k = (3,3), s = 1, p = 1$	$1024 \times 14 \times 14$
Reshape2		$256 \times 28 \times 28$
Conv5	$k = (1,1), s = 1, p = 0$	$2 \times 28 \times 28$
Output		$2 \times 28 \times 28$

Table 1: The parameters for each block in our proposed mask generator. k, s, and p denote the kernel size, stride, and padding of the convolution operation, respectively.

ure 4 . The details of the parameters of each block are shown in Table 1.

Feature Fusion

As shown in Figure 3, after obtaining the mask predictions M and bounding box predictions B, the feature map F_0 from FPN is fused together with them to derive a new semantic segmentation feature map F_N, where N is the total number of mask and bounding box predictions, $i \in 1, ..., N$. For the ith bounding box prediction, the coordinates of the bounding box with the highest class score are selected, which is defined as:

$$B_i = (x_i, y_i, w_i, h_i) \tag{1}$$

where x_i and y_i are the coordinates of the top left point of B_i in x and y axes, respectively, w_i and h_i are its width and height, respectively. When fusing the ith mask M_i with B_i on the feature map F_{i-1}, the output feature map F_i is formulated as:

$$F_i = G_{replace}(G_{reshape}(M_i, B_i), sub(F_{i-1}, B_i)) \tag{2}$$

where $G_{reshape}$ reshapes the binary mask M_i to the same size as B_i. $G_{replace}$ enables the new mask $G_{reshape}(M_i, B_i)$ to replace $sub(F_{i-1}, B_i)$, which is the subset of F_{i-1} according to B_i and can be represented as:

$$sub(F_{i-1}, B_i) = F_{i-1}[x_i : x_i + w_i, y_i : y_i + h_i] \tag{3}$$

Eventually, F_N is the semantic feature map obtained from the instance branch.

For each ROI, mask prediction represents the nuclei segmentation result in the small region by utilizing the local information around each single object. In addition, the bounding box prediction represents the real location of this nucleus in the original image, which contains localization information for the object in a global view. In order to retain more global and local information in the semantic segmentation result, we fuse the mask and bounding box predictions by passing each mask prediction feature map to the large feature map from FPN according to its corresponding box coordinates prediction. Therefore, this second semantic segmentation feature map (F_N) contains global and local features from the detection architecture of the instance branch.

3.2 Semantic Branch

Even though a semantic feature map F_N is rendered from the instance branch, the small kernel sizes of the convolutional layers in FPN imply that F_N would still have some information loss at the global level. This is a common issue for CNN models in that as the network grows deeper, the actual receptive field of the feature maps gradually becomes smaller than the theoretical size [Zhou *et al.*, 2015]. By utilizing small sizes convolutional layers [Saleh *et al.*, 2018], it is difficult for the feature maps at high resolution to maintain the whole features from original images due to the limit of the receptive field.

To tackle this issue, the decoder of Global Convolutional Network (GCN) [Peng *et al.*, 2017] is applied as an auxiliary semantic branch shown in Figure 2. By simulating a 2D large kernel convolutional layer with two 1D convolutional layers, GCN is able to make the model capture a large and global receptive field with only a small portion of memory. Our model follows the original GCN architecture except the size of the large kernel blocks which, for memory efficiency, is fixed at 5. In addition, the semantic branch is trained jointly with the instance branch, by sharing the same backbone, compared to the traditional separate training strategy in [Saleh *et al.*, 2018; Kirillov *et al.*, 2018]. When working on the semantic segmentation task, the backbone is capable of generating the features with global information about foreground and background, which are useful for bounding box detection and mask generation in the instance branch.

3.3 Loss Function

All the losses in this work are shown in Figure 2. The total loss is defined as:

$$\begin{aligned} L_{total} = L_{(rpn-obj)} + L_{(rpn-reg)} + L_{(det-cls)} \\ + L_{(det-reg)} + L_{(det-mask)} \\ + L_{semseg1} + L_{semseg2} \end{aligned} \tag{4}$$

$L_{(rpn-obj)}$ and $L_{(rpn-reg)}$ are the losses for background and foreground classification and the bounding boxes for the anchors rendered by RPN, respectively, where $L_{(rpn-obj)}$ is the cross entropy loss for classification and $L_{(rpn-reg)}$ is the smoothed L1 loss for regression. In the instance branch, $L_{(det-cls)}$ is the cross entropy loss for object category classification, $L_{(det-reg)}$ is the smoothed L1 loss for bounding box coordinate regression, and $L_{(det-mask)}$ is the binary cross entropy loss for instance mask segmentation. $L_{semseg1}$ and $L_{semseg2}$ are the cross entropy losses for semantic segmentation of semantic and instance branch, respectively. Although adding a weight for each loss in L_{total} would be a better trade-off, we still chose to avoid this labor-intensive process, in favor of the generalizability and reproducibility.

In all experiments, we employed stochastic gradient descent (SGD) as the optimizer with a momentum of 0.9 and a weight decay of 0.0001 to train our model. The learning rate varies in each experiment with the same linear warming up in the first 500 iterations. Due to the small mini-batch size during training, we had no batch normalization layers. All the hyperparameters for testing were fine-tuned on the validation set. We implemented our experiments using Pytorch [Paszke *et al.*, 2017].

4 Experiments

4.1 Datasets and Evaluation Metrics

We used three public datasets in this study. The first dataset is from The Cancer Genome Atlas (TCGA) at $40\times$ magnification [Kumar *et al.*, 2017]. We refer to this dataset as TCGA-kumar. There are 30 annotated 1000×1000 patches from 30 whole slide images of different patients. These images show highly varying properties since they are from 18 different hospitals and 7 different organs (breast, liver, kidney, prostate, bladder, colon, and stomach). In contrast to the disease variability of TCGA-kumar, the second dataset from [Naylor *et al.*, 2018] focuses in particular on Triple Negative Breast Cancer (TNBC). In this TNBC dataset, there are 50 annotated 512×512 patches at $40\times$ magnification sampled from 11 patients at the Curie Institute. The third dataset is the MICCAI 2017 Digital Pathology Challenge dataset [Vu *et al.*, 2018], also referred to as Cell17. Cell17 contains 64 annotated images in total, and both the training and testing sets contain 8 images from 4 different diseases: glioblastoma multiforme (GBM), lower grade glioma (LGG) tumors, head and neck squamous cell carcinoma (HNSCC), and non small cell lung cancer (NSCLC). The image sizes are either 500×500 or 600×600 at $20\times$ or $40\times$ magnification.

In the experiments on TCGA-kumar and TNBC, we employed F1 score and Aggregated Jaccard Index (AJI) [Kumar *et al.*, 2017] for pixel-level and object-level evaluation, respectively. AJI is an extension of the Jaccard Index that takes false negative predictions into consideration. AJI can be defined as:

$$AJI = \frac{\sum_{i=1}^{N} |G_i \cap P_M^i|}{\sum_{i=1}^{N} |G_i \cup P_M^i| + \sum_{F \in U} |P_F|} \tag{5}$$

where G_i is the ith nucleus from the ground truth with N nuclei. P_M^i means the Mth connected component in prediction which has the largest Jaccard Index with G_i, and each index (M) cannot be used more than once. Set U represents the connected component in the prediction without the corresponding ground truth. In the experiments of Cell17, we employed the same set of evaluation metrics as the experiments of Cell R-CNN [Zhang *et al.*, 2018a] for comparison: F1 score, object-level Dice score, and object-level Hausdorff distance.

4.2 Experimental Results and Discussion

TCGA-kumar

In this experiment, we first evaluated the performance of our proposed model in comparison to the state-of-the-art works from [Kumar *et al.*, 2017; Naylor *et al.*, 2018; He *et al.*, 2017; Ronneberger *et al.*, 2015]. Then, an ablation study was employed to demonstrate the effectiveness of each module in the overall architecture.

With the same split for the testing set as [Kumar *et al.*, 2017] and [Naylor *et al.*, 2018] (details at https://peterjacknaylor.github.io/), we compared our work directly to the results reported in their published works. For [He *et al.*, 2017], we re-implemented it with officially released code from https://github.com/facebookresearch/maskrcnn-benchmark. Among the 16 training images from four different organs, we randomly selected one image from each organ for validation and used the remaining 12 images for training. The learning rate in this experiment was 0.0025 and it decreased to its $1/10$ at the $8640th$ iteration and $1/100$ at the $12960th$ iteration with a total of 17280 training iterations. When training the model, we separated each 1000×1000 original image into four 512×512 patches with basic data augmentation including horizontal and vertical flipping and rotations of $90°$, $180°$, and $270°$ to avoid overfitting. In addition, we also employed advanced augmentation including blurring, adding gaussian noise, embossing, sharpening, and random channel shuffle due to the noise and variability of color in the histopathology images. For a fair comparison, we applied the same data augmentation settings to all compared methods.

Table 2 shows the result for each image in the testing set in comparison to 2 state-of-the-art nuclei segmentation methods: CNN3 in [Kumar *et al.*, 2017] and DIST in [Naylor *et al.*, 2018]. Our proposed method outperforms the others based on the average AJI and F1 score. In addition, one-tailed paired t-test was employed for evaluation of statistically significance. For object-level accuracy (AJI), the improvement of our method is significant as all the p-values of the comparison methods are under 0.1. By learning the global and local features from both the semantic and instance branches, our model is the most competitive even without any pre- or post-processing compared to CNN3 and DIST. In terms of F1 score, our improvement is significant compared to the others except for DIST. However, DIST has a post-processing step with two extra hyperparameters, in order to refine the distance map from a deep regression model. If a less significant improvement is acceptable, we would prefer not to add any further post-processing for the sake of memory efficiency and convenience of implementation. For Mask R-CNN, we notice that its AJI score is at the same level as CNN3, while the F1 score is lower than all the compared semantic segmentation models. This is because Mask R-CNN processes ROI, which represents each object for instance segmentation, making it impossible to learn the relationship between the foreground and background. By adding a new semantic branch, an extra semantic loss, and a dual-modal mask generator, our proposed model has a significant improvement in pixel-level accuracy.

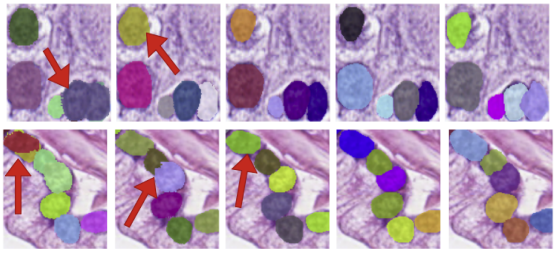

Figure 5: Visual comparison of the ablation study on the TCGA-kumar dataset. The first row and second row show results of images from the breast and prostate, respectively. The red arrows indicate the disagreement with the ground truth.

Organ	Image	AJI					F1 Score				
		CNN3	DIST	Mask R-CNN	U-Net	Proposed	CNN3	DIST	Mask R-CNN	U-Net	Proposed
Breast	1	0.4974	0.5334	0.4960	0.3831	0.5490	0.6885	0.7761	0.7486	0.7492	0.7887
	2	0.5796	0.5884	0.5577	0.5505	0.6362	0.7476	0.8380	0.8026	0.8144	0.8478
Kidney	1	0.4792	0.5648	0.5794	0.5386	0.6202	0.6606	0.7805	0.7742	0.8039	0.8097
	2	0.6672	0.5420	0.5286	0.5020	0.5902	0.7837	0.7606	0.7285	0.7786	0.7724
Liver	1	0.5175	0.5466	0.5183	0.4773	0.5491	0.6726	0.7877	0.7816	0.7648	0.7907
	2	0.5148	0.4432	0.4577	0.3794	0.5179	0.7036	0.6684	0.6881	0.6313	0.7323
Prostate	1	0.4914	0.6273	0.5934	0.1807	0.6305	0.8306	0.8030	0.8032	0.7889	0.8090
	2	0.3761	0.6294	0.6282	0.3118	0.6423	0.7537	0.7903	0.7937	0.7919	0.7992
Bladder	1	0.5465	0.6475	0.6237	0.5115	0.6749	0.9312	0.8623	0.8385	0.8258	0.8666
	2	0.4968	0.5467	0.4677	0.4621	0.4745	0.6304	0.7768	0.6944	0.7648	0.6963
Colon	1	0.4891	0.4240	0.3691	0.0786	0.4450	0.7679	0.7212	0.6251	0.7121	0.7036
	2	0.5692	0.4484	0.4354	0.1305	0.4871	0.7118	0.7360	0.6907	0.7599	0.7474
Stomach	1	0.4538	0.6408	0.6352	0.4096	0.6909	0.8913	0.8547	0.8323	0.8647	0.8795
	2	0.4378	0.6550	0.6449	0.4507	0.6871	0.8982	0.8520	0.8329	0.8629	0.8668
Average		0.5083	0.5598	0.5382	0.3833	**0.5854**	0.7623	0.7863	0.7596	0.7795	**0.7936**
Significance		⋆⋆	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆		⋆	✗	⋆⋆⋆	⋆	

Table 2: Comparison with the state-of-the-art methods on TCGA-kumar dataset. For statistical significance evaluation, ⋆⋆⋆ denotes p-value under 0.01, ⋆⋆ denotes p-value from 0.01 to 0.05, ⋆ denotes p-value from 0.05 to 0.1, and ✗ denotes p-value over 0.1

Name		Baseline	Ab1	Ab3	Ab2	Proposed
feature fusion?		✗	✓	✗	✓	✓
semantic branch?		✗	✗	✓	✓	✓
dual-modal mask generator?		✗	✗	✗	✗	✓
AJI	avg	0.5382	0.5533	0.5500	0.5744	0.5854
	std	0.0851	0.0770	0.0912	0.0782	0.0820
	significance		⋆⋆	⋆⋆	⋆⋆⋆	⋆⋆
F1	avg	0.7596	0.7730	0.7670	0.7881	0.7936
	std	0.0655	0.0572	0.0717	0.0545	0.0591
	significance		⋆⋆	⋆	⋆⋆⋆	⋆

Table 3: Ablation study on the TCGA-kumar dataset. ⋆⋆⋆ is employed if the p-value is under 0.01, ⋆⋆ for p-value from 0.01 to 0.05, ⋆ for p-value from 0.05 to 0.1. The significance of Ab1, Ab3, and Proposed are comparisons between Ab1 and Baseline, Ab3 and Baseline, Proposed and Ab2, respectively. The significance of Ab2 are the comparisons between Ab2 and Ab1, Ab2 and Ab3, and both p-values are under 0.01.

In order to evaluate the effectiveness of each proposed module in our architecture, an ablation experiment was conducted and the results are shown in Figure 5 and Table 3. Figure 5 shows that the baseline Mask R-CNN tends to fail in segmenting some touching or closely adjacent nuclei and Ab1 (baseline + feature fusion) has difficulty in accurately estimating the boundary information. Compared with Ab2 (baseline + feature fusion + semantic branch), our method has a higher accuracy when generating the mask for single nucleus in detail. In these experiments, we used the same settings and data as the comparison experiment. In addition to comparing the average and standard deviation, we also calculated the p-value between the architecture with and without each module using one-tailed paired t-test. As shown in Table 3, all the improvements after adding each module are significant ($p < 0.1$). Compared with Ab3 (baseline + semantic branch), Ab2 has a higher accuracy, which indicates the feature fusion module is more effective than semantic branch on TCGA-kumar dataset.

TNBC

We were interested in whether our model with two branches works better than the model with a single semantic or instance branch. For this, we designed an experiment on the TNBC dataset by comparing our proposed model with Mask R-CNN [He *et al.*, 2017] and GCN [Peng *et al.*, 2017]. In addition, we also compared with pixel2pixel [Isola *et al.*, 2017] to prove the effectiveness of our model without any adversarial based techniques. In this experiment, we either used the code from official implementation or re-implemented them in Pytorch. For each experiment, we used six cases with 30 images for training, two cases with seven images for validation, and three cases with 13 images for testing. For data augmentation, we employed horizontal and vertical flipping and rotations of $90°$, $180°$, and $270°$. The total training epoch was 60 and the initial learning rate 0.00075 decreased to its $1/10$ and $1/100$ at the end of the $30th$ and $45th$ epoch, respectively.

Method		GCN	Pix2Pix	MRCNN	Proposed
AJI	avg	0.1907	0.4760	0.5297	0.5865
	std	0.1208	0.0578	0.1513	0.1059
	significance	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	
F1	avg	0.3833	0.6910	0.7424	0.7792
	std	0.2175	0.0724	0.0837	0.0520
	significance	⋆⋆⋆	⋆⋆⋆	⋆⋆⋆	

Table 4: The result on the TNBC testing set for different methods. ⋆⋆⋆ denotes p-value under 0.01, ⋆⋆ if the p-value between 0.01 and 0.05, ⋆ if p-value between 0.05 and 0.1.

Table 4 lists the resulting AJI and F1 score on the testing set. Our proposed model outperformed all the comparison models in both average AJI and F1 score. In addition, one-tailed paired t-test was also employed to analyze whether our improvements were statistically significant compared to the others. The p-values of all the comparison methods are noted in Table 4. For both F1 and AJI score, all the improvements of our method are significant ($p < 0.1$). For the semantic segmentation model, only class labels were assigned to each pixel and it was unable to separate different instance objects within the same category. By generating ROIs for predicting the location and mask for each single instance object with the same class label, Mask R-CNN has a higher accuracy at both the pixel and object level compared to the semantic segmentation models. However, due to the nature of the Mask R-CNN architecture, only the object-level information is taken into account, which makes it difficult for the model to process the semantic information. By incorporating the semantic branch with the instance branch, our model is capable of processing the global information from the semantic branch.

Method	F1-score	Dice	Hausdorff
Pix2Pix	0.6208 ± 0.1126	0.6351 ± 0.0706	19.1441 ± 6.0933
FnsNet	0.7413 ± 0.0668	0.6165 ± 0.0839	25.9102 ± 9.5834
Mask R-CNN	0.8004 ± 0.0722	0.7070 ± 0.0598	12.6723 ± 3.4591
Cell R-CNN	0.8216 ± 0.0625	0.7088 ± 0.0564	11.3141 ± 2.6917
Proposed	$\mathbf{0.8645 \pm 0.0482}$	$\mathbf{0.7506 \pm 0.0491}$	$\mathbf{9.5832 \pm 3.6237}$

Table 5: The quantitative results for the Cell17 dataset.

Cell17

We designed Cell17 experiment to compare our proposed model to Cell R-CNN from [Zhang *et al.*, 2018a], which is also a panoptic segmentation architecture for nuclei segmentation and demonstrates the state-of-the-art performance in a recent study. In addition, classic semantic and instance segmentation models include Pix2Pix [Isola *et al.*, 2017], FnsNet [Johnson *et al.*, 2016], and Mask R-CNN [He *et al.*, 2017] are compared as well. Among 32 training images from four different organs, our validation set contained one image randomly selected from each organ, while the remaining 28 images were used for training. In this experiment, we employed basic data augmentation including horizontal and vertical flipping and rotations of $90°$, $180°$, and $270°$. The total training epoch was 100, and the initial learning rate was 0.001, decreasing to $1/10$ and $1/100$ of the initial learning rate at the end of the $50th$ and $75th$ epochs, respectively. All the comparison results are from [Zhang *et al.*, 2018a].

As shown in Table 5, our proposed work outperforms all the compared models in all the three metrics. We noticed that the object-level Dice score of Cell R-CNN is at the same level as Mask R-CNN. This is because only the instance branch was used during inference, which makes the prediction tend to rely on fewer global features from the foreground and background. To address the problem, we added a new semantic loss from the instance branch in this work so that the instance branch is able to learn the relationship between the things and stuff as well.

Discussion

For previous CNN based nuclei segmentation methods, the semantic segmentation models were able to generate results with high accuracy at the pixel level. However, the object level accuracy was limited due to the inability to process local information inside and around the nuclei. Recently, region based CNN such as Mask R-CNN have been prevalent for instance segmentation by processing the ROIs which contain the features for each object. Although the result of Mask R-CNN has a high object level accuracy, failing to process global information results in poor pixel level accuracy. In order to address this problem, we incorporate both semantic and instance segmentation branches.

In a larger perspective, the segmentation tasks for medical image are not limited to nuclei segmentation for histopathology images. With the significant improvement of pixel- and object-level accuracy in the experiments of this work, we hope that our proposed architecture will contribute to other medical, or even ordinary imaging applications.

5 Conclusion

In this work, we propose a panoptic segmentation architecture for nuclei segmentation in histopathology images. By jointly training the semantic branch with large convolutional kernels and instance segmentation branches with a feature fusion mechanism, our model is able to incorporate the complementary information at both global and local levels. Results of extensive nuclei segmentation experiments on three public datasets indicate that our method is highly effective and outperforms all the compared methods by a large margin.

Acknowledgments

We gratefully acknowledge funding provided by the following National Institutes of Health (NIH) grants: P41 EB015902, P41 EB015898, R01 MH074794, R01 MH119222, and U01 CA199459.

References

[Chankong *et al.*, 2014] Thanatip Chankong, Nipon Theera-Umpon, and Sansanee Auephanwiriyakul. Automatic cervical cell segmentation and classification in pap smears. *Computer Methods and Programs in Biomedicine*, 113(2):539–556, 2014.

[Chen *et al.*, 2017] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. DCAN: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis*, 36:135–146, 2017.

[de Geus *et al.*, 2018] Daan de Geus, Panagiotis Meletis, and Gijs Dubbelman. Panoptic segmentation with a joint semantic and instance segmentation network. *arXiv preprint arXiv:1809.02110*, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988. IEEE, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[Kirillov *et al.*, 2018] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv preprint arXiv:1801.00868*, 2018.

[Kirillov *et al.*, 2019] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. *arXiv preprint arXiv:1901.02446*, 2019.

[Kong *et al.*, 2013] Hui Kong, Hatice Cinar Akakin, and Sanjay E Sarma. A generalized laplacian of gaussian filter

for blob detection and its applications. *IEEE Transactions on Cybernetics*, 43(6):1719–1733, 2013.

[Kumar *et al.*, 2017] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE Transactions on Medical Imaging*, 36(7):1550–1560, 2017.

[Liu *et al.*, 2018a] Dongnan Liu, Donghao Zhang, Siqi Liu, Yang Song, Haozhe Jia, Dagan Feng, Yong Xia, and Weidong Cai. Densely connected large kernel convolutional network for semantic membrane segmentation in microscopy images. In *ICIP*, pages 2461–2465. IEEE, 2018.

[Liu *et al.*, 2018b] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, pages 8759–8768. IEEE, 2018.

[Naik *et al.*, 2008] Shivang Naik, Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *ISBI*, pages 284–287. IEEE, 2008.

[Naylor *et al.*, 2018] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. Segmentation of nuclei in histopathology images by deep regression of the distance map. *IEEE Transactions on Medical Imaging*, 2018.

[Oda *et al.*, 2018] Hirohisa Oda, Holger R Roth, Kosuke Chiba, Jure Sokolić, Takayuki Kitasaka, Masahiro Oda, Akinari Hinoki, Hiroo Uchida, Julia A Schnabel, and Kensaku Mori. Besnet: Boundary-enhanced segmentation of cells in histopathological images. In *MICCAI*, pages 228–236. Springer, 2018.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[Peng *et al.*, 2017] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters—improve semantic segmentation by global convolutional network. In *CVPR*, pages 1743–1751. IEEE, 2017.

[Raza *et al.*, 2019] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical image analysis*, 52:160–173, 2019.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Saleh *et al.*, 2018] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *ECCV*, pages 86–103. Springer, 2018.

[Veta *et al.*, 2014] Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014.

[Vu *et al.*, 2018] Quoc Dang Vu, Simon Graham, Minh Nguyen Nhat To, Muhammad Shaban, Talha Qaiser, Navid Alemi Koohbanani, Syed Ali Khurram, Tahsin Kurc, Keyvan Farahani, Tianhao Zhao, et al. Methods for segmentation and classification of digital microscopy tissue images. *arXiv preprint arXiv:1810.13230*, 2018.

[Wang *et al.*, 2018] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *WACV*, pages 1451–1460. IEEE, 2018.

[Wang *et al.*, 2019] Heng Wang, Donghao Zhang, Yang Song, Siqi Liu, Dagan Feng, Yue Wang, Hanchuan Peng, and Weidong Cai. Segmenting neuronal structure in 3d optical microscope images via knowledge distillation with teacher-student network. In *ISBI*, pages 228–231. IEEE, 2019.

[Wienert *et al.*, 2012] Stephan Wienert, Daniel Heim, Kai Saeger, Albrecht Stenzinger, Michael Beil, Peter Hufnagl, Manfred Dietel, Carsten Denkert, and Frederick Klauschen. Detection and segmentation of cell nuclei in virtual microscopy images: a minimum-model approach. *Scientific Reports*, 2:503, 2012.

[Xiong *et al.*, 2019] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. *arXiv preprint arXiv:1901.03784*, 2019.

[Zhang *et al.*, 2014] Ling Zhang, Hui Kong, Chien Ting Chin, Shaoxiong Liu, Zhi Chen, Tianfu Wang, and Siping Chen. Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts. *Computerized Medical Imaging and Graphics*, 38(5):369–380, 2014.

[Zhang *et al.*, 2018a] Donghao Zhang, Yang Song, Dongnan Liu, Haozhe Jia, Siqi Liu, Yong Xia, Heng Huang, and Weidong Cai. Panoptic segmentation with an end-to-end cell R-CNN for pathology image analysis. In *MICCAI*, pages 237–244. Springer, 2018.

[Zhang *et al.*, 2018b] Donghao Zhang, Yang Song, Siqi Liu, Dagan Feng, Yue Wang, and Weidong Cai. Nuclei instance segmentation with dual contour-enhanced adversarial network. In *ISBI*, pages 409–412. IEEE, 2018.

[Zhang *et al.*, 2018c] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, pages 269–284, 2018.

[Zhou *et al.*, 2015] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *ICLR*, 2015.