Panoptic Feature Fusion Net: A Novel Instance Segmentation Paradigm for Biomedical and Biological Images

Article *in* IEEE Transactions on Image Processing · January 2021

DOI: 10.1109/TIP.2021.3050668

Dongnan Liu
The University of Sydney

Donghao Zhang
The University of Sydney

Yang Song
UNSW Sydney

Weidong Cai
The University of Sydney

Panoptic Feature Fusion Net: A Novel Instance Segmentation Paradigm for Biomedical and Biological Images

Dongnan Liu, Donghao Zhang, Yang Song, *Member, IEEE*, Heng Huang, and Weidong Cai, *Member, IEEE*

***Abstract*— Instance segmentation is an important task for biomedical and biological image analysis. Due to the complicated background components, the high variability of object appearances, numerous overlapping objects, and ambiguous object boundaries, this task still remains challenging. Recently, deep learning based methods have been widely employed to solve these problems and can be categorized into proposal-free and proposal-based methods. However, both proposal-free and proposal-based methods suffer from information loss, as they focus on either global-level semantic or local-level instance features. To tackle this issue, we present a Panoptic Feature Fusion Net (PFFNet) that unifies the semantic and instance features in this work. Specifically, our proposed PFFNet contains a residual attention feature fusion mechanism to incorporate the instance prediction with the semantic features, in order to facilitate the semantic contextual information learning in the instance branch. Then, a mask quality sub-branch is designed to align the confidence score of each object with the quality of the mask prediction. Furthermore, a consistency regularization mechanism is designed between the semantic segmentation tasks in the semantic and instance branches, for the robust learning of both tasks. Extensive experiments demonstrate the effectiveness of our proposed PFFNet, which outperforms several state-of-the-art methods on various biomedical and biological datasets.**

***Index Terms*— Instance segmentation, panoptic segmentation, histopathology images, fluorescence microscopy images, plant phenotype images.**

I. Introduction

INSTANCE segmentation is a prerequisite step for biomedical and biological image processing, which not only assigns a class label for each pixel but also separates each object within the same class. By assigning a unique ID for every single object, the morphology, spatial locations, and distribution of the objects can be further studied to analyze the biological behaviors from the given images. In the digital pathology domain, the nuclear pleomorphism (size and shape) contributes to the tumor and cancer grading, and the spatial arrangement of cancer nuclei facilitates the understanding of cancer prognostic predictions [1]–[4]. In the plant and agriculture study, analyzing each distinguished leaf in plant images enables the experts to learn about the plant phenotype including the number of leaves, maturity condition, and its similar cultivars, which serves as the key factor of understanding plant function and growth condition [5]–[7]. Traditional manual assessment for biomedical and biological image instance segmentation is not suitable for current practice, as it is labor-intensive and time-consuming. Additionally, limitations of objective and reproducibility are unavoidable due to the intra- and inter-observer variability [8]. To this end, automatic and accurate methods for instance segmentation in biology images are necessary and in high demand.

There still remain some challenges in instance segmentation tasks for biomedical and biological images. First, some background structures have a similar appearance to the foreground object, such as cytoplasm or stroma in histopathology images. Therefore, methods relying on thresholding are ineffective. Second, within the same dataset, the objects in different images have large variability in size, shape, texture, and intensity. It is caused by the various biological structures and activities when acquiring different images [9], [10]. Third, there are clusters of objects overlapping with each other. The boundaries between these touching objects are ambiguous due to nonuniform staining absorption and similar object intensity. This might result in segmenting several objects into a single one. In order to tackle these issues, deep learning based methods are prevalent and effective by learning from feature representations.

CNN based instance segmentation methods can be categorized into two types: proposal-free and proposal-based methods. For the proposal-free instance segmentation methods, each pixel is firstly assigned a class label with a semantic segmentation model. The post-processing steps are then employed to separate each foreground object within the same category, according to their morphology characteristic, structures, and spatial arrangement [11]–[15]. Although post-processing among these methods is capable of separating the connected components, they still suffer from artificial boundaries during overlapping object segmentation. Even though [11], [15], [16] focus on boundaries learning at the semantic segmentation stage, the global contextual information is still not enough to separate the touching objects, especially when their borders become unclear. On the other hand, the proposal-based

Manuscript received February 17, 2020; revised November 2, 2020; accepted December 30, 2020. Date of publication January 15, 2021; date of current version January 22, 2021. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Christophoros Nikou. *(Corresponding author: Dongnan Liu.)*

Dongnan Liu, Donghao Zhang, and Weidong Cai are with the School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia (e-mail: dliu5812@uni.sydney.edu.au).

Yang Song is with the School of Computer Science and Engineering, University of New South Wales Sydney, Sydney, NSW 2052, Australia.

Heng Huang is with the Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA 15261 USA, and also with the JD Finance America Corporation, Mountain View, CA 94043 USA.

Digital Object Identifier 10.1109/TIP.2021.3050668

instance segmentation methods incorporate the detection task with the segmentation task [17], [18]. First, the spatial location for each object is detected as a bounding box. Then, a mask generator is further employed to segment each object within the corresponding predicted bounding box. By detecting and segmenting every single object separately, the proposal-based methods are capable of separating the touching objects. However, they are limited as there is a lack of global semantic information between the foreground and background.

For the instance segmentation tasks, both the global semantic and local instance information is important. The global semantic information indicates the useful clues in the scene context, such as the relationship between the foreground and background and the spatial distribution of all the foreground objects. On the other hand, local-level instance information describes the spatial location and detailed contour for every single object. To integrate the benefits of the global and local features, panoptic segmentation [19], reconciliation of the semantic and instance segmentation, has been proposed. In [19], the predictions from two separately trained semantic and instance segmentation branches are fused together to analyze the panoptic level segmentation. Without sharing components between the two branches, training [19] incurs a large computational cost [20]. In addition, the analysis in [20] indicates that jointly training a network for the two tasks achieves better performance than training them independently. To this end, Panoptic PFN [20] is proposed to jointly train the semantic and instance segmentation branches by sharing the same ResNet backbone, which has achieved state-of-the-art performance on panoptic segmentation as well as maintained memory efficiency.

Based on [19] and [20], we previously proposed [21] and [22] for nuclei instance segmentation in histopathology images. Motivated by jointly analyzing the semantic and instance segmentation tasks in [19], we designed the Cell R-CNN [21] to simultaneously process the global and local information in the histopathology images. Different from the two separately optimized branches in [19], our Cell R-CNN [21] proposed to jointly train the two branches with a shared backbone model. In order to further facilitate the semantic-level contextual learning in the instance segmentation model, our IJCAI work [22] was proposed to induce the instance branch to learn directly about the semantic-level features. As the extension of Cell R-CNN, we refer to [22] as Cell R-CNN V2 in the following sections. In Cell R-CNN V2, we firstly introduce a new semantic segmentation prediction from the instance branch. Then a feature fusion mechanism to incorporate the feature maps is designed to induce the semantic feature learning in the decoder of the instance segmentation branch, by integrating the mask prediction of the instance branch with that of the semantic branch. In addition, a dual-model mask generator is proposed for instance mask segmentation, in order to prevent information loss. Compared with the Panoptic FPN [20], which only jointly optimized the semantic and instance segmentation branches with a shared backbone, our Cell R-CNN V2 directly integrated the features from the two branches, to further induce the semantic feature learning in the instance branch.

In this work, we propose a Panoptic Feature Fusion Net (PFFNet), which further extends our preliminary Cell R-CNN V2 [22] by addressing several remaining problems. First, the feature fusion mechanism in [22] directly replaces the part of the feature map in the semantic segmentation branch with those from the output of the mask generator. Although the mask predictions from the instance branch interpret more instance-level features than the semantic branch, the global contextual features from the semantic segmentation prediction are also important. To this end, we propose a residual attention feature fusion mechanism (RAFF) in this work, to replace the previous feature fusion mechanism. In our newly proposed RAFF, the local features from the instance branch are integrated with the global semantic features, without deprecating any semantic-level features. Second, two semantic segmentation tasks with the same ground truth are optimized together in the overall architecture of [22]. In order to facilitate the robust learning of two segmentation tasks, we add a semantic consistency regularization between them to enforce the two semantic predictions from two different branches as similar as possible. In addition, there remain some low-quality mask predictions with an unexpected high classification score in the traditional Mask R-CNN, as mentioned in [23]. It would be harmful to the segmentation accuracy if treating these poorly generated results as the ones with high confidence. To this end, we propose a new mask quality sub-branch in this work, by learning an auxiliary quality score of each mask prediction based on the Dice score and Intersection-over-Union (IoU) score. During inference, the classification score of each mask is re-weighted through multiplication by its corresponding mask quality score.

The PFFNet proposed in this manuscript is an extension of Cell R-CNN V2, and can therefore also be named Cell R-CNN V3. In line with our previous Cell R-CNN V2 and Cell R-CNN, we are the first to employ the panoptic segmentation idea on biomedical and biological image analysis, to the best of our knowledge. Overall, the contributions of this work compared with Cell R-CNN V2 are summarized as follows:

- We design a residual attention feature fusion mechanism to integrate the features of each detected object in the semantic and instance levels.
- We design a semantic task consistency mechanism to regularize the semantic segmentation tasks training for robustness.
- We design an extra mask quality sub-branch to ensure the mask segmentation quality for each object is compatible with its confidence score.
- Our proposed Panoptic Feature Fusion Net is validated on the instance segmentation tasks for various biomedical and biological datasets, including histopathology images, fluorescence microscopy images, and plant phenotyping images. Our results for all metrics outperform the state-of-the-art methods by a large margin.

II. Related Work

Instance segmentation for biomedical and biological images is widely studied, ranging from the handcrafted feature-based methods to the learning-based methods. In order to emphasize

the contributions of our proposed PFFNet, we mainly focus on the literature of deep learning based instance segmentation methods, which can be grouped into two classes: the proposal-free and proposal-based methods.

A. Proposal-Free Instance Segmentation

Proposal-free instance segmentation methods are mainly based on the morphology and spatial relationship of all the objects in the images. For example, object boundary is an important feature for separating the touching object. In [11], [15], [16], [24], the instances are separated according to the probability map for the foreground objects and their boundaries. Similarly, [25] separates each instance according to the distance between the two connected components. Additionally, post-processing methods are employed to separate the touching objects based on the semantic segmentation predictions, such as conditional region growing algorithm [11], morphological dynamics algorithms [11], and watershed algorithm [15], [16]. In addition to the traditional classification-based segmentation methods, regression-based methods are also widely employed. In [12], a distance transform map describing the distance between each pixel and its nearest background pixel is predicted, with a regression CNN architecture. To obtain the instance segmentation map directly, [13], [14], [26] employ the clustering algorithm on the high dimensional embeddings predicted from the deep regression CNN model. Based on adversarial learning architecture, Zhang et al [16] proposed an image-to-image translation method for a more accurate probability map compared with the classification-based method.

B. Proposal-Based Instance Segmentation

Compared with proposal-free instance segmentation methods, the proposal-based methods predict the mask segmentation for each object based on the predictions of their corresponding locations in the whole image [27], [28]. One fundamental proposal-based instance segmentation method is Mask R-CNN [17]. Based on the high-dimensional feature maps from the backbone CNN network, Mask R-CNN firstly generates regions of interest (ROIs) containing the foreground objects with a region proposal network (RPN). After aligning the ROIs to the same size, a box sub-branch and a mask sub-branch are employed to predict the coordinate, class label, and mask prediction for each ROI. With the help of the local-level information from the spatial locations of the instances, Mask R-CNN achieved state-of-the-art performance compared with the traditional box-free methods. Following the Mask R-CNN, other methods were further proposed with a higher accuracy: [18] proposed a path aggregate backbone to preserve the feature maps at high resolutions, [23] added a branch for mask IoU score prediction based on the mask prediction on the original Mask R-CNN, and [29] employed a cascade connection of several bounding box and mask prediction sub-branches.

Although the proposal-based instance segmentation methods achieve higher performance compared with the proposal-free methods by processing each object separately, their effectiveness is still limited due to the lack of the semantic-level global information on the context of the whole images. To tackle this issue, panoptic segmentation was recently proposed to jointly process the foreground things and the background stuff [19], by incorporating the semantic segmentation with the instance segmentation. Inspired by this joint segmentation idea, [30] fuses the instance segmentation result for foreground objects with the semantic segmentation result for the background for urban scene semantic segmentation. However, the instance branch and semantic branch are trained separately in previous work. In [20], [31], both the instance and semantic segmentation branches are trained together by sharing the same backbone module. Then, the losses of the two branches are summed together for back propagation to optimize the parameters of the whole framework. Later, more methods for fusing the results of things and stuff are proposed. In [32], attention mechanism is employed to fuse the proposals and masks from the instance branch with the feature map from the semantic branch. [33] proposed a spatial ranking module to separate the overlapping objects from different categories by fusing the semantic segmentation predictions with the instance segmentation ones.

Similar to the jointly learning paradigm in the panoptic segmentation, combining the semantic segmentation task of the proposal-based instance segmentation also enables the model to achieve higher performance by learning the auxiliary semantic-level contextual information. In [34], the semantic segmentation prediction is fused with the proposed hybrid cascade instance segmentation architecture to make the architecture manipulate the global semantic features and achieve state-of-the-art performance compared with previous instance segmentation methods. In medical analysis tasks, we previously proposed Cell R-CNN [21] to induce the encoder of the instance segmentation to learn semantic-level information by jointly training a semantic segmentation network and a Mask R-CNN with a shared backbone network. With the help of the semantic-level contextual information, Cell R-CNN outperforms Mask R-CNN in the nuclei segmentation tasks on histopathology images. However, the decoder of the Cell R-CNN only learns the semantic features indirectly, which still makes the model lack global information during inference. In Cell R-CNN V2 [22], we, therefore, designed a feature fusion module to incorporate the feature maps from the semantic segmentation branch and the instance segmentation branch during the training phase. By retaining semantic-level features in the encoder and decoder of the instance segmentation model, our previous work [22] achieved state-of-the-art performance on several nuclei instance segmentation tasks under both object- and pixel-level metrics.

III. Panoptic Feature Fusion Net

In this section, we firstly introduce the overall architecture of the proposed Panoptic Feature Fusion Net (PFFNet). Then, the three newly proposed modules are described in detail. Finally, the training and inference details are presented.

A. Overall Architecture

Fig. 1 illustrates our proposed PFFNet. For each input image, it first passes through a ResNet-101 [35] backbone

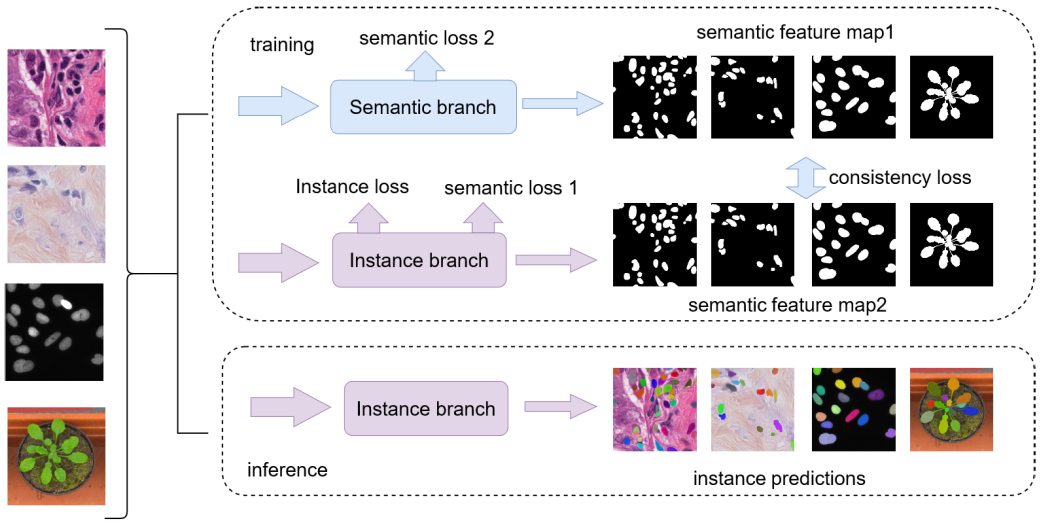

Fig. 1. Overview of our proposed Panoptic Feature Fusion Net (PFFNet). The input images first passed through a backbone network for multi-resolution feature maps. The backbone is omitted for brevity. The overall loss function for training is shown in Eq.5.

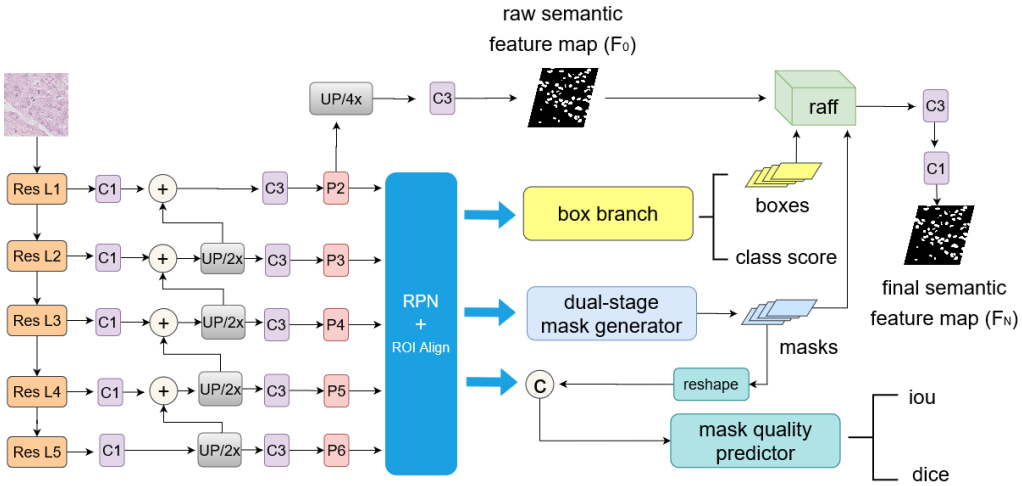

Fig. 2. Overview of our instance segmentation branch. $C1$ and $C3$ represent the convolutional layer with the kernel size of 1 and 3 respectively. The C before the mask quality predictor is the concatenation operation. UP/nx means the upsampling layer for n times with the nearest interpolation. $raff$ represents the proposed residual attention feature fusion mechanism. The ReLU and group normalization layer after all the convolutional layers are omitted for brevity.

network to obtain the feature maps at different resolutions. Then, the feature maps are sent to a semantic segmentation branch to learn the global semantic-level feature and an instance segmentation branch to learn the object-level local features.

For the semantic segmentation branch, we employed the decoder of the global convolutional network (GCN) [36], as shown in Fig. 3. Specifically, multi-resolution feature maps after the ResNet101 backbone network are sent to a skip connected decoder, which contains several large kernel global convolutional modules. Each large kernel global convolutional module is simulated by incorporating two 1D convolutional kernels in different orders. To this end, the model has a large receptive field as well as memory efficiency, and the semantic branch is capable of processing more global-level contextual features compared with the CNN architectures with a normal size convolutional kernels.

Our instance segmentation branch in Fig. 2 is based on that of Cell R-CNN V2 [22]. First, multi-resolution feature maps ($P2$, $P3$, $P4$, $P5$, and $P6$ in Fig. 2) are obtained by the feature pyramid network (FPN) [37] connected after the backbone encoder. Along with the anchors in different ratios and sizes, $P2$, $P3$, $P4$, $P5$, and $P6$ then pass through a region proposal network (RPN) [38] to generate ROIs which represent the features of all possible foreground objects in the original images. As the ROIs after RPN are in various sizes, a ROIAlign mechanism [17] is further employed to reshape all the ROIs to the same size, which is 14×14 in this work. Eventually, all the ROIs are sent to a bounding box sub-branch to predict the locations and class scores and a dual-model mask generator [22] for mask instance segmentation prediction. In order to induce the semantic feature learning in the decoder of the instance segmentation branch, we further propose an attention-based feature fusion mechanism to incorporate the

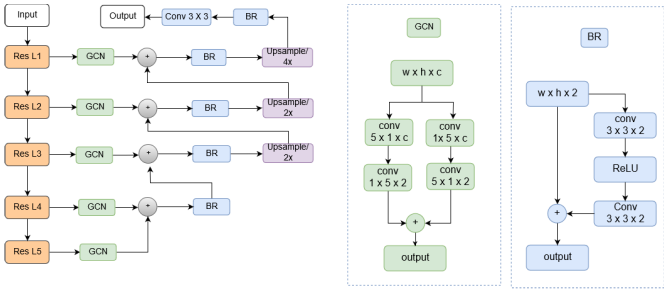

Fig. 3. Overview of the semantic branch in our proposed PFFNet.

mask prediction and bounding box prediction for all the ROIs with the semantic segmentation feature map obtained from the top layer of FPN ($P2$). In addition, the mask segmentation result is fused with the ROI features for a newly proposed mask quality sub-branch to predict the quality of the mask segmentation for each ROI according to the corresponding IoU and Dice score.

B. Residual Attention Feature Fusion Mechanism

In Cell R-CNN V2, we proposed a feature fusion mechanism to incorporate the semantic-level contextual features with the local-level instance features by using the mask prediction from the instance branch to replace the subset of the semantic segmentation features according to the location of the bounding box sub-branch. Although the fused feature map contains both semantic- and instance-level features, only the background features at the semantic level are learned by the instance segmentation branch, as the foreground features in the original semantic feature map are deprecated. However, the foreground feature for each object from the global view in the original semantic feature maps is still important, as it contains the relationship between each object and the whole background. In the instance segmentation branch, the mask prediction of each object is predicted according to the relationship between the foreground and the background within the corresponding 28×28 ROI, instead of the background of the whole image. Moreover, part of the background feature in the semantic feature map is also replaced by that from the instance predictions after the feature fusion mechanism in the Cell R-CNN V2, which results in the contextual information loss in the semantic segmentation prediction. To this end, directly replacing the subset of the semantic feature map with the predictions of the instance branch is harmful to the semantic-level feature learning in the decoder of the instance branch.

To tackle this issue, we design an attention-based feature fusion mechanism, as illustrated in Fig. 4. The number of ROIs in the instance segmentation branch is denoted as K, and the mask and bounding box predictions for each ROI are defined as M_i and B_i, respectively, where $i \in [1, K]$. Specifically, B_i can be written as:

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

where x_i and y_i represent the corrdinates of the bottom left point of the ith rectangle ROI in x and y axes and w_i and h_i are its width and height. In addition, the semantic feature map before the attention-based feature fusion is defined as F_0, as illustrated in Fig. 2. During the attention-based feature fusion for each M_i, first we obtain its probability map P_i:

$$P_i = \sigma(M_i) \tag{2}$$

where $\sigma()$ is the sigmoid operation. Then, we fuse each P_i with the subset of F_0 according to the correpsonding B_i. as shown in Algorithm 1, where $R(P_i, (w_i, h_i))$ reshapes the P_i to (w_i, h_i) with bilinear interpolation, and $*$ is the element-wise multiplication.

Algorithm 1 Algorithm for the Residual Attention Feature Fusion Mechanism

Require:
Mask probability predictions in the instance branch P_i, its corresponding bounding box B_i, and the raw semantic feature F_0. $i = 1, \ldots, K$

for $i \in [1, K]$ **do**
  $sub(F_0, B_i) = F_{i-1}[x_i : x_i + w_i, y_i : y_i + h_i]$;
  $F_i = sub(F_0, B_i) * (1 + R(P_i, (w_i, h_i)))$;
end for
return F_K

The value of each coordinate of P_i represents the probability of this pixel being the foreground. Therefore, the proposed residual attention feature fusion mechanism highlights the foreground features on the original semantic feature map. By fusing the instance-level features on the semantic feature map while preserving all its contextual features, optimizing the semantic segmentation task of the instance branch enables the mask generator to learn accurate and sufficient semantic features.

C. Mask Quality Sub-Branch

During the inference process of the traditional Mask R-CNN, the mask predictions are determined by the highest classification score. However, the classification scores for the mask predictions are not always correlated with their quality, such as the IoU between the mask prediction and the ground truth [23]. In the testing phase of the Cell R-CNN and Cell R-CNN V2, if there remain two overlapping predictions, the overlapping part is assigned to the mask with a higher classification score. Therefore, low-quality mask predictions with high classification scores affect the performance when processing the overlapping objects during inference.

Inspired by [23], we propose a new mask quality sub-branch to predict the quality of the mask predictions in the instance branch, as shown in Fig. 2. For each mask prediction in size $2 \times 28 \times 28$, we select its foreground $1 \times 28 \times 28$ score map. Then, each $1 \times 28 \times 28$ score map is reshaped to size $4 \times 14 \times 14$, to concatenate with the $256 \times 14 \times 14$ ROI feature map. The fused feature map with size $260 \times 14 \times 14$ then passes through 3 convolutional layers and 3 fully connected layers to predict the quality of the mask, which is a float value in (0, 1). Table I indicates the detailed hyperparameters setting in the mask quality sub-branch. As Dice coefficient is an important

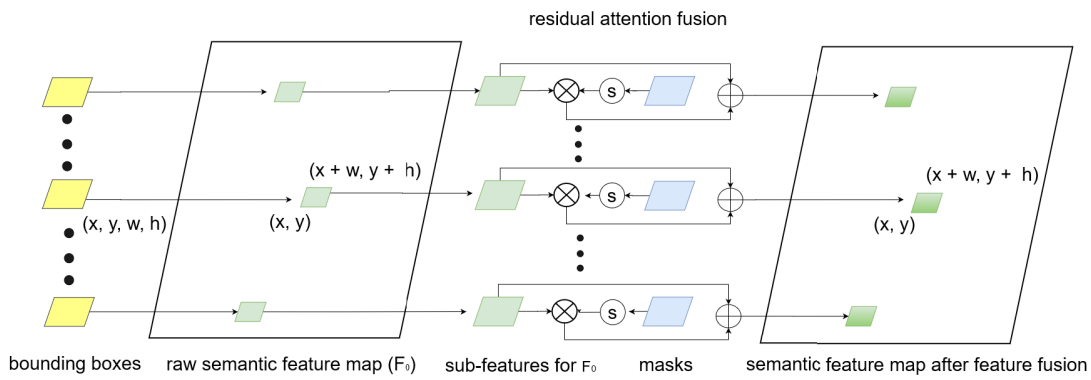

Fig. 4. The proposed residual attention feature fusion mechanism. S is the sigmoid operation, $\times$ is the element-wise multiplication, and $+$ is the element-wise summation.

TABLE I
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

Stage	Hyperparamaters	Output size
Input		$260 \times 14 \times 14$
Conv1	$k = (3, 3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv2	$k = (3, 3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv3	$k = (3, 3), s = 1, p = 1$	$256 \times 14 \times 14$
Conv4	$k = (3, 3), s = 2, p = 1$	$256 \times 7 \times 7$
FC1		$1 \times 1 \times 1024$
FC2		$1 \times 1 \times 1024$
FC3		$1 \times 1 \times 1$
Output		$1 \times 1 \times 1$

metric in the biomedical and biological images segmentation task, the mask quality score is determined by the IoU and Dice score between the predictions and the ground truth. During training, a mask prediction and its corresponding ground truth are denoted as M_p and M_t, respectively, the mask quality score s_{qua} is defined as:

$$s_{qua} = (2 * \frac{|M_p \cap M_t|}{|M_p| + |M_t|} + \frac{|M_p \cap M_t|}{|M_p \cup M_t|}) * 0.5 \tag{3}$$

where $|.|$ means the total number of the pixels. Therefore, s_{qua} is in (0, 1). Eventually, $l2$ loss is employed between the s_{qua} and the mask quality prediction.

D. Semantic Task Consistency Regularization

Our motivation for this module is from [39]. When there are two tasks in a multi-task learning architecture focusing on the same objective, adding a consistency regularization between the outputs of these two tasks enables the robust learning of both. In our proposed architecture, both the semantic and the instance branches generate semantic segmentation predictions. In the ideal situation, the semantic segmentation predictions from both two branches should be equal to each other and equal to the ground truth. Therefore, we propose a consistency regularization between these two semantic segmentation predictions to reduce the distance between them. The softmax semantic segmentation prediction of the semantic and instance branch are denoted as p_{sem} and p_{ins}, respectively, which are both in range $(0, 1)$. The semantic consistency regularization is:

$$L_{sem-cons} = \frac{1}{N} \sum_{i,j} (p_{sem(i,j)} - p_{ins(i,j)})^2 \tag{4}$$

where N is the total number of activations in the p_{sem}.

E. Training and Inference Details

As shown in Fig. 1, the total loss function of the PFFNet is defined as:

$$\begin{aligned} L_{overall} = & L_{rpn-obj} + L_{rpn-reg} + L_{det-cls} \\ & + L_{det-reg} + L_{det-mask} + L_{det-qua} \\ & + \alpha_1 (L_{semseg1} + L_{semseg2}) + \alpha_2 L_{sem-cons} \end{aligned} \tag{5}$$

For the instance segmentation task, $L_{rpn-obj}$ and $L_{rpn-cls}$ are the smooth L1 regression loss and cross entropy classification loss for RPN, respectively. $L_{det-reg}$ and $L_{det-cls}$ are the bounding box regression and the classification loss of the box sub-branch, $L_{det-mask}$ is the binary cross entropy segmentation loss for the mask sub-branch, and $L_{det-qua}$ is the $l2$ regression loss for the mask quality sub-branch. On the other hand, $L_{semseg1}$ and $L_{semseg2}$ are the semantic segmentation losses for the semantic branch and instance branch. $L_{sem-cons}$ is the mean square loss for the semantic consistency regularization, as shown in Eq. 4. α_1 and α_2 are trade-off parameters to balance the importance of each task and are set as 0.1 and 1, respectively, in our experiments.

During inference, the instance mask predictions from the mask generator of the instance branch are employed. A confident threshold score β is firstly employed to depreciate the masks whose classification scores are smaller than β. Then, a mask confidence score s_{conf} for each object is calculated based on its classification score s_{cls} and mask quality prediction s_{qua}:

$$s_{conf} = s_{cls} * s_{qua} \tag{6}$$

For any two touching predictions, the overlapping part belongs to the prediction with the higher s_{conf}.

IV. Experiments

A. Dataset Description

1) TCGA-KUMAR: This dataset contains 30 histopathology images in size 1000 × 1000, obtained from the The Cancer Genome Atlas (TCGA) at 40× magnification [11]. Each image is from one of the seven organs, including breast, bladder, colon, kidney, liver, prostate, and stomach. In order to compare with the state-of-the-art methods, we have the same data split as in [11], [12], [22]. 12 images total from the breast, kidney, liver, and prostate are employed for training (3 from each organ). During training, 20 patches in size 256 × 256 are randomly cropped from each 1000 × 1000 image. Next, basic augmentation techniques are applied, including horizontal and vertical flipping and rotation of 90°, 180°, and 270°. Due to the noise and variability of color in the histopathology images, advanced augmentation including Gaussian blur, median blur, Gaussian noise are then employed to ensure the robustness of the model. The validation set contains 4 images from the breast, kidney, liver, and prostate. For the remaining 14 images, 8 images from the same 4 organs in the training set form the seen testing set, while 6 from the other 3 organs unavailable to the training are selected as the unseen testing set. During testing, each 1000×1000 image is directly employed for nuclei instance segmentation.

2) TNBC: This is our second histopathology dataset focusing on the Triple Negative Breast Cancer (TNBC) dataset from [12]. The TNBC dataset contains 30 512 × 512 histopathology images at 40× magnification, collected from 11 different patients of the Curie Institute. We conduct 3-fold cross validation for all the experiments on this dataset. During training, 5 256 × 256 patches are cropped from each 512 × 512 images, following data augmentation including including horizontal and vertical flipping, rotation of 90°, 180°, and 270°, Gaussian blur, median blur, and Gaussian noise. For testing, each 512 × 512 image is directly employed.

3) Fluorescence Microscopy Images: In addition to the histopathology images, we also validate our PFFNet on the fluorescence microscopy images analysis. We employ the BBBC039V1 dataset from [40], which contains 200 520 × 696 images obtained from fluorescence microscopy. Each image focuses on the U2OS cells with a single field of view on the DNA channel, with various cell shape and density. In our experiment, we follow the official data split (https://data.broadinstitute.org/bbbc/BBBC039/), with 100 images for training, 50 for validation, and the rest 50 for testing. For training data preparation, first, 10 256 × 256 patches are randomly cropped from each image. As the background components in this dataset are not as complicated as the others, only basic data augmentation is employed, including horizontal and vertical flipping and rotation of 90°, 180°, and 270°. During inference, each 520 × 696 image is directly used.

4) Plant Phenotyping: To demonstrate the effectiveness of our proposed PFFNet on instance segmentation task for other biology images, we study the leaf instance segmentation task. We employ the Computer Vision Problems in Plants Phenotyping (CVPPP) [5] dataset, which contains top-down view images of leaves with various shapes and complicated occlusions. In this work, we focus on the A1 subset with a total of 161 530 × 500 images, which has been broadly studied for instance segmentation in several state-of-the-art works. Out of the 128 training images provided by the challenge, we employed 100 images for training and the remaining 28 for validation. During training, each image is firstly reshaped to size 512 × 512. Then, data augmentation including horizontal and vertical flipping, rotation of 90°, 180°, and 270°, Gaussian blur, median blur, and Gaussian noise are employed to avoid overfitting. During inference, the predictions are directly obtained from the 530 × 500 images. To evaluate the performance, the predicted results are submitted to the official evaluation platform (https://competitions.codalab.org/competitions/18405).

B. Evaluation Metrics

To evaluate the performance on the nuclei segmentation in the histopathology images and cell segmentation in the fluorescence microscopy images, we employed Aggregated Jaccard Index (AJI), object-level F1 score ($F1$), Panoptic Quality (PQ), and pixel-level Dice score ($Dice$). AJI is an extended Jaccard Index for object-level segmentation evaluation [11], defined as:

$$AJI = \frac{\sum_{i=1}^{N} |G_i \cap P_M^i|}{\sum_{i=1}^{N} |G_i \cup P_M^i| + \sum_{F \in U} |P_F|} \tag{7}$$

where G_i is the ith nucleus in a ground truth with a total of N nuclei. U is the set of false positive predictions without the corresponding ground truth. For each ground truth object G_i, M is the index of the prediction with the largest overlapping with it and each M can only be used once, which is defined as:

$$M = argmax \frac{P_M^i \cap G_i}{P_M^i \cup G_i} \tag{8}$$

Object-level F1 score is the metric for the detection performance [15], defined based on the number of true and false detections:

$$F1 = \frac{2TP}{FN + 2TP + FP}, \tag{9}$$

where TP, FN, and FP represent the number of true positive (corrected detected objects), false negative (ignored objects), and false positive (detected objects without corresponding ground truth) detections, respectively. Note that a true positive object for object-level F1 score should intersect with more than 50% of its corresponding ground truth. Panoptic Quality (PQ) has been previously employed to evaluate the performance of the panoptic segmentation tasks [19], [20], which is the multiplication between the Detection Quality (DQ) for object detection, and Segmentation Quality (SQ) for object segmentation. PQ is defined as:

$$PQ = \underbrace{\frac{2|TP|}{2|TP| + |FP| + |FN|}}_{DQ} \times \underbrace{\frac{\sum_{(p,g) \in TP} IoU(p,g)}{|TP|}}_{SQ} \tag{10}$$

where $|TP|$, $|FN|$, and $|FP|$ represent the number of true positive, false negative, and false positive detections, respectively. Each (p, g) indicates a pair of mask prediction from the

TABLE II

THE COMPARISON OF RESULTS FOR TCGA-KUMAR DATASET. AVG AND STD REPRESENT AVERAGE AND STANDARD DEVIATION, RESPECTIVELY. FOR DIST, THE RESULTS OF OBJECT-LEVEL $F1$ AND PQ ARE UNKNOWN. FOR CNN3, THE PQ SCORE IS UNKNOWN

Methods		AJI			$Dice$			$F1$			PQ		
		seen	unseen	all	seen	unseen	all	seen	unseen	all	seen	unseen	all
CNN3 [11]	avg	0.5154	0.4989	0.5083	0.7301	0.8051	0.7623	0.8226	0.8322	0.8267	–	–	–
	std	0.0835	0.0806	0.0695	0.0590	0.1006	0.0946	0.0853	0.0764	0.0934	–	–	–
DIST [12]	avg	0.5594	0.5604	0.5598	0.7756	0.8005	0.7863	–	–	–	–	–	–
	std	0.0598	0.0663	0.0781	0.0489	0.0538	0.0550	–	–	–	–	–	–
Mask R-CNN [17]	avg	0.5438	0.5340	0.5396	0.7659	0.7658	0.7659	0.6987	0.6434	0.6750	0.4856	0.4715	0.4796
	std	0.0649	0.1283	0.0929	0.0481	0.0608	0.0517	0.1344	0.1908	0.1566	0.0893	0.1709	0.1248
Cell R-CNN [21]	avg	0.5547	0.5606	0.5572	0.7746	0.7752	0.7748	0.7587	0.7481	0.7542	0.5066	0.5098	0.5079
	std	0.0567	0.1100	0.0800	0.0446	0.0577	0.0485	0.0969	0.1488	0.1166	0.0816	0.1392	0.1051
Cell R-CNN V2 [22]	avg	0.5758	0.5999	0.5861	0.7841	0.8078	0.7943	0.8014	0.8023	0.8017	0.5500	0.5563	0.5527
	std	0.0568	0.1160	0.0841	0.0439	0.0611	0.0512	0.0757	0.1081	0.0871	0.0748	0.1346	0.1000
PFFNet	avg	**0.5975**	**0.6282**	**0.6107**	**0.7967**	**0.8256**	**0.8091**	**0.8317**	**0.8383**	**0.8345**	**0.5824**	**0.5933**	**0.5871**
	std	0.0568	0.0924	0.0726	0.0453	0.0520	0.0487	0.0694	0.0598	0.0631	0.0767	0.1023	0.0850

true positive detections, and its corresponding ground truth. Note that a mask prediction can only be regarded as the true positive when $IoU(p, q) > 0.5$. As illustrated in Eq. 10, the PQ metric reflects the performance on object detection and segmentation. To evaluate the foreground and background segmentation accuracy, pixel-level Dice score is employed between the binarized prediction and the ground truth:

$$Dice = \frac{2|P \cap G|}{|P| + |G|} \tag{11}$$

where P and G represent the binarization prediction and ground truth, respectively. $|.|$ means the total number of foreground pixels.

For the evaluation metrics of the leaf segmentation task, we directly employ the official Symmetric Best Dice (SBD) score:

$$SBD(P, T) = min(BD(P, T), BD(T, P)) \tag{12}$$

where P and T are the predictions and ground truth, respectively. $BD(P, T)$ is the best dice between $P_{i(}i = 1, \ldots, M)$ and $T_{j(}j = 1, \ldots, N)$:

$$BD(P, T) = \frac{1}{M} \sum_{i=1}^{M} \max_{j=1,\ldots,N} \frac{2|P_i \cap T_j|}{|P_i| + |T_j|} \tag{13}$$

where $|.|$ means the total number of foreground pixels.

C. Implementation Details

For the network initialization, the weights of the ResNet101 backbone are pretrained on the ImageNet [41] classification task, while the weights for other layers are initialized with "Kaiming" initialization [42]. When training the PFFNet, stochastic gradient descent (SGD) is used to optimize the network, with a weight decay of 0.0001, and momentum of 0.9. The mini-batch size is 1, which is relatively a small batch size. We, therefore, employed group normalization layers [43] with a group number of 32 to replace the traditional batch normalization layers. The initial learning rate is set to 0.003, with a linear warm-up for the first 500 iterations. The learning rate is then decreased to 0.0003 when it reaches the $3/4$ of the total training iterations. Our experiments are implemented on two Nvidia GeForce 1080Ti GPUs with Pytorch [44].

TABLE III

P-VALUE FOR THE METHODS IN TABLE II COMPARED WITH OUR PROPOSED PFFNET, ON TCGA KUMAR DATASET UNDER ALL THE FOUR METRICS

	AJI	$Dice$	$F1$	PQ
CNN3	2.69×10^{-3}	1.67×10^{-2}	0.34	–
DIST	2.35×10^{-6}	1.27×10^{-5}	–	–
Mask R-CNN	6.44×10^{-5}	1.48×10^{-6}	5.57×10^{-5}	1.14×10^{-5}
Cell R-CNN	3.93×10^{-6}	2.39×10^{-6}	2.53×10^{-4}	6.14×10^{-7}
Cell R-CNN V2	7.02×10^{-4}	1.03×10^{-5}	1.33×10^{-3}	1.07×10^{-4}

D. Comparison With State-of-the-Art Instance Segmentation Methods for Biomedical and Biological Images

1) TCGA-KUMAR: Our result is compared with several state-of-the-art nuclei instance segmentation methods, including CNN3 [11], DIST [12], Mask R-CNN [17], Cell R-CNN [21], and Cell R-CNN V2 [22]. With the same data split, we directly compare the performance reported in [11], [12]. For Mask R-CNN, Cell R-CNN, and Cell R-CNN V2, we re-implement them by adding group normalization with the same settings as our proposed PFFNet, for a fair comparison. Therefore their performance is slightly better than in [22]. Table II and Fig. 5 illustrate our quantitative and qualitative comparison results, respectively. As shown in Table II, our proposed PFFNet outperforms all the other methods in all four metrics on the seen and unseen testing set. It indicates that our PFFNet has a strong generalization ability when testing on the cases from the unseen organs. In order to test the statistical significance between the results of our PFFNet and other methods, we employed one-tailed-paired t-test to calculate the p-value. As shown in Table III, our improvements under all four metrics is statistically significant (p-value < 0.05) except for the $F1$ of CNN3. However, $F1$ only relies on the number of corrected detected objects, regardless of the segmentation quality of each detected object. By outperforming CNN3 by a large margin in the other two segmentation metrics (over 10% on AJI and 4% on $Dice$), our PFFNet still achieves better performance on nuclei segmentation tasks compared with CNN3. Fig. 6 is the box plot for all the compared method under the four metrics, which shows that our proposed PFFNet not only outperforms all the methods, but is also more stable and robust.

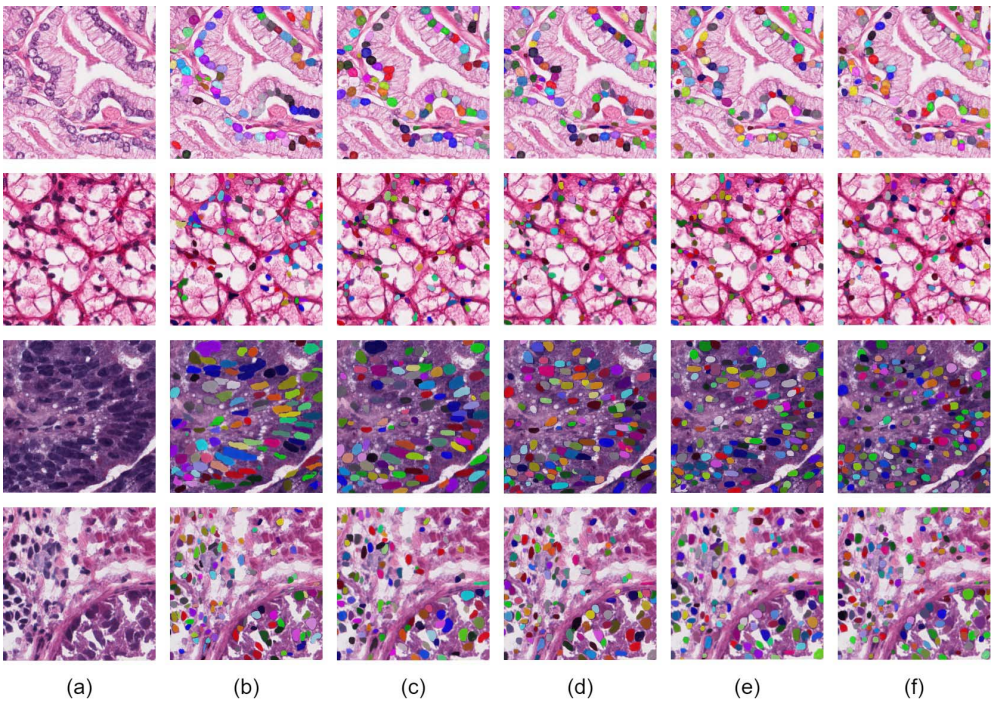

Fig. 5. The visual comparison of results for TCGA KUMAR dataset. (a) original images, (b) ground truth annotations, (c) predictions by our proposed PFFNet, (d) predictions by Cell R-CNN V2 [22], (e) predictions by Cell R-CNN [21], and (f) predictions by Mask R-CNN [17].

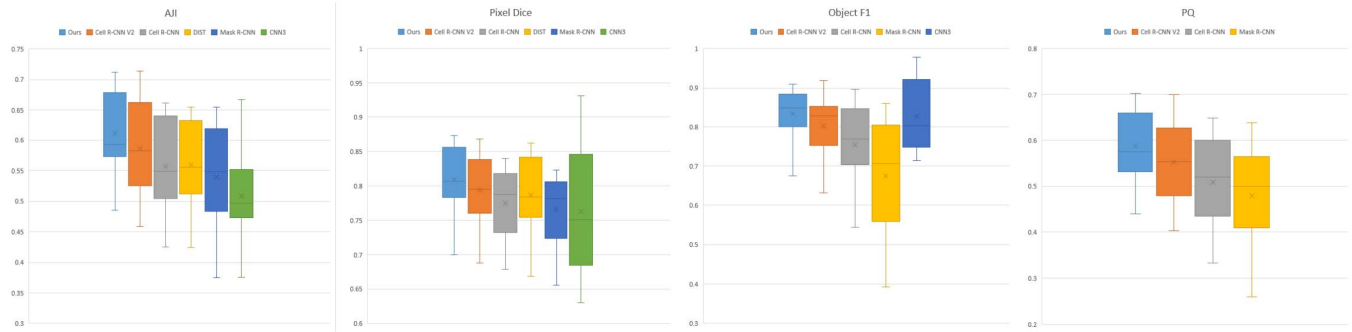

Fig. 6. Box plot for all the compared methods on TCGA KUMAR dataset. The object-level F1 and PQ score and were not available in the original DIST work. The PQ score was not reported by the original CNN3 work.

TABLE IV

COMPARISON EXPERIMENTS ON TNBC DATASET UNDER AJI, $Dice$, $F1$, AND PQ. RESULTS ARE PRESENTED AS MEAN VALUE WITH STANDARD DEVIATION IN THE PARENTHESES

Methods	AJI	$Dice$	$F1$	PQ
Mask R-CNN	0.5350(0.0993)	0.7393(0.0977)	0.7542(0.1535)	0.5146(0.1193)
Cell R-CNN	0.5747(0.1061)	0.7637(0.1080)	0.8142(0.1331)	0.5664(0.1120)
Cell R-CNN V2	0.5986(0.0847)	0.7793(0.0772)	0.8184(0.1163)	0.5845(0.0964)
PFFNet	**0.6313 (0.0750)**	**0.8037 (0.0557)**	**0.8600 (0.0849)**	**0.6298 (0.0820)**

2) TNBC: We conducted comparison experiments on the second histopathology dataset with 3-fold cross-validation and the results are shown in Table IV and Fig. 7. As in Table IV, our PFFNet outperforms its previous versions under all three metrics. Compared with Mask R-CNN, the effectiveness of the Cell R-CNN is improved by a large margin. The background components in the TNBC dataset are complicated and some background textures have a similar appearance to the foreground. Therefore, processing the semantic-level information is beneficial to the segmentation and detection accuracies. Compared with the Cell R-CNN, the improvement of Cell R-CNN V2 is not as large as in the TCGA KUMAR dataset, especially under the object-level F1 score. Although the feature fusion mechanism in the Cell R-CNN V2 facilitates the semantic feature learning in the instance branch, there is a lack of contextual features around

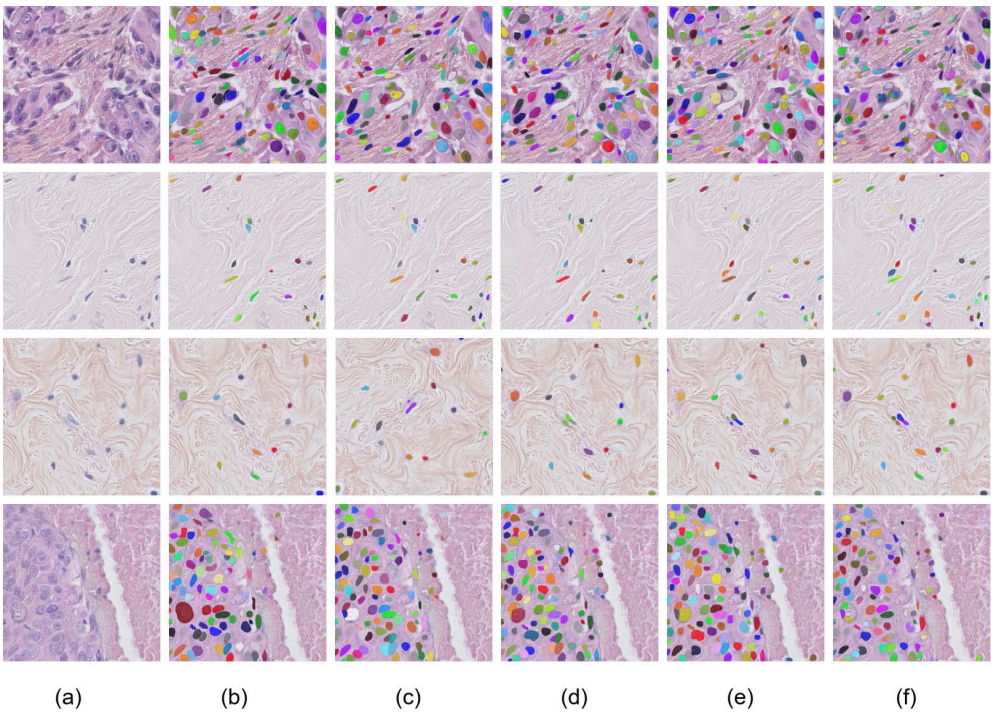

Fig. 7. The visual comparison of results for TNBC dataset. (a) original images, (b) ground truth annotations, (c) predictions by our proposed PFFNet, (d) predictions by Cell R-CNN V2 [22], (e) predictions by Cell R-CNN [21], and (f) predictions by Mask R-CNN [17].

TABLE V

COMPARISON EXPERIMENTS ON BBBC039V1 DATASET UNDER AJI, $Dice$, $F1$, AND PQ. RESULTS ARE PRESENTED AS MEAN VALUE WITH STANDARD DEVIATION IN THE PARENTHESES

Methods	AJI	$Dice$	$F1$	PQ
Mask R-CNN	0.7983(0.0858)	0.9277(0.0126)	0.9180(0.0870)	0.7773(0.0959)
Cell R-CNN	0.8070(0.0934)	0.9290(0.0273)	0.9276(0.0836)	0.7959(0.0894)
Cell R-CNN V2	0.8260(0.0779)	0.9336(0.0097)	0.9328(0.0728)	0.8010(0.0839)
PFFNet	**0.8477 (0.0757)**	**0.9478 (0.0071)**	**0.9451 (0.0536)**	**0.8331 (0.0724)**

each object due to the depreciation of part of the semantic feature map. Therefore, the detection accuracies are affected when the boundaries of two touching objects become ambiguous. Similar to the results on the TCGA KUMAR dataset, our proposed PFFNet outperforms the compared methods by a large margin.

3) BBBC039V1: In addition to the nuclei segmentation tasks in the histopathology images, our proposed PFFNet is also effective for cell instance segmentation in the fluorescence microscopy images. As illustrated in Table V, our PFFNet outperforms all the compared methods. We notice that the performance of Cell R-CNN is at the same level as Mask R-CNN, due to the limited improvement. In the fluorescence microscopy images, the background components are not as complicated as in the histopathology images, as shown in Fig. 8. Therefore, Cell R-CNN is not capable of improving accuracy as it fails to process the contextual information about the foreground objects by learning the semantic features in the backbone encoder. On the other hand, Cell R-CNN V2 improves Cell R-CNN by designing a dual-modal mask generator for improving the mask segmentation accuracies in the instance branch and inducing the mask generator to learn global semantic-level features. However, the improvement of the pixel-level Dice score of Cell R-CNN is still limited. Based on Cell R-CNN V2, our proposed PFFNet achieves high improvements in all four metrics.

TABLE VI

COMPARISON EXPERIMENTS ON CVPPP DATASET WITH OUR PREVIOUS METHODS. RESULTS ARE PRESENTED AS MEAN VALUE WITH STANDARD DEVIATION IN THE PARENTHESES

Methods	SBD	$Dice$	PQ
Mask R-CNN	0.8674(0.0362)	0.8453(0.0799)	0.6342(0.0860)
Cell R-CNN	0.8771(0.0274)	0.8772(0.0524)	0.6804(0.0750)
Cell R-CNN V2	0.8861(0.0259)	0.8931(0.0434)	0.7161(0.0669)
PFFNet	**0.9062 (0.0269)**	**0.9338 (0.0268)**	**0.7788(0.0552)**

4) CVPPP Challenge: First, we compare with our previous work, with the 3-fold cross-validation on the 128 training images. The results are shown in Table VI and Fig. 9. By outperforming our previous Cell R-CNN and Cell R-CNN V2 on the instance segmentation tasks for biology images as well as the medical images, our proposed PFFNet is demonstrated to be effective.

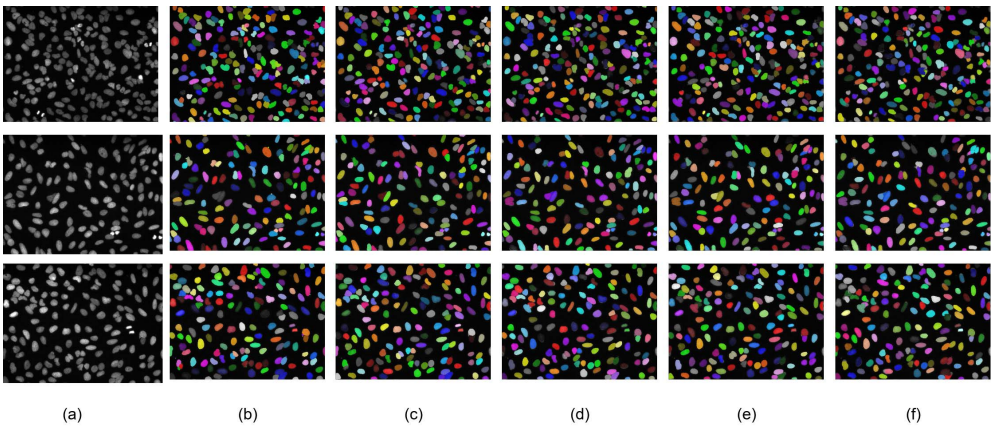

Fig. 8. The visual comparison of results for BBBC039 V1 dataset. (a) original images, (b) ground truth annotations, (c) predictions by our proposed PFFNet, (d) predictions by Cell R-CNN V2 [22], (e) predictions by Cell R-CNN [21], and (f) predictions by Mask R-CNN [17].

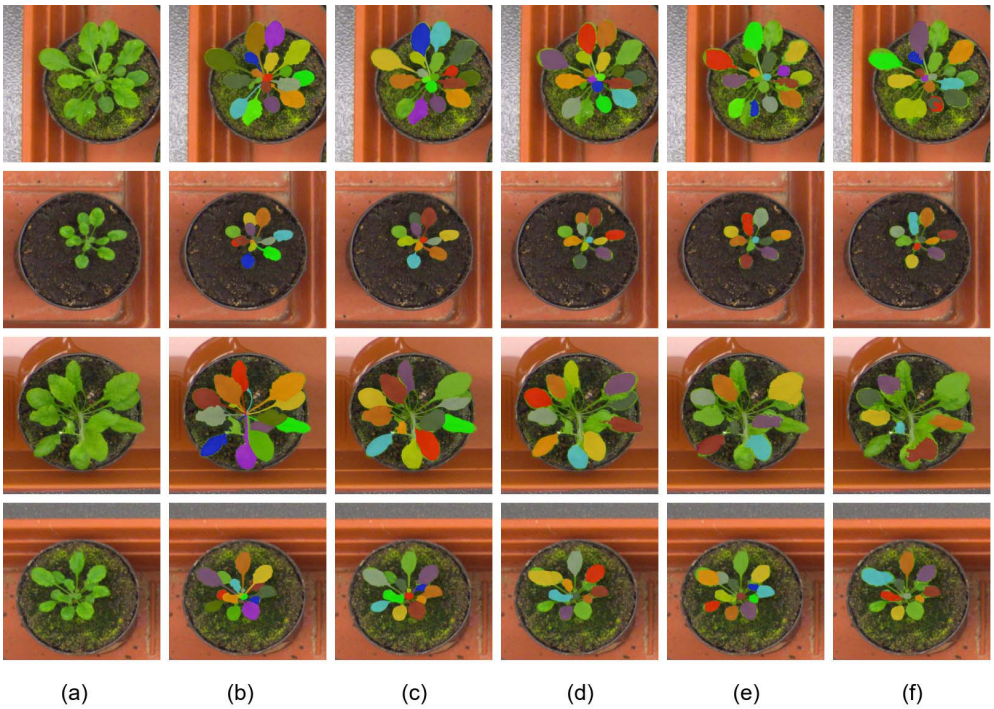

Fig. 9. The visual comparison of results for CVPPP A1 dataset. (a) original images, (b) ground truth annotations, (c) predictions by our proposed PFFNet, (d) predictions by Cell R-CNN V2 [22], (e) predictions by Cell R-CNN [21], and (f) predictions by Mask R-CNN [17].

To further demonstrate the effectiveness of our PFFNet on biology image analysis, we also conducted a comparison experiment with other previous work, using the 33 leaf segmentation testing images. Table VII is the performance between our work and the state-of-the-art methods and our segmentation accuracy outperforms all the existing published work on this dataset. Among these methods, RIS [45], RNN [46], and Recurrent with attention [27] processed one instance each time, with the help of the temporal chain from recurrent neural work (RNN) or long-short-term memory (LSTM). In addition, [27] achieved better performance compared with the previous [45] and [46] due to the attention module and proposal-based architecture. As there is actually no temporal information in the leaf instance segmentation task, other methods focusing on the spatial relationship are more suitable for the task with better performance. [10], [14], [25], and [26] are proposal-free instance segmentation methods. [25] divides all the objects into several groups of untouching instances and processes them separately. Instead of directly learning the instance mask prediction for each leaf, [10], [14], and [26] learned high-dimensional embedding maps projected from the original images. Then, clustering algorithms were employed to separate each instance during inference. Without focusing on each object, their performance is still limited due to the lack of local-level information. Similar to our work, the CNN architecture in [47] and [48] are proposal-based Mask R-CNN. With the help of the auxiliary synthesized images, these two methods outperform most of

TABLE VII
THE QUANTITATIVE LEAF SEGMENTATION RESULTS FOR THE CVPPP A1 DATASET COMPARED WITH THE STATE-OF-THE-ART METHODS

Method	SBD (%)
RIS [45]	66.6
RNN [46]	74.7
Deep coloring [25]	80.4
Embedding-based [13]	84.2
Discriminative loss [14]	84.2
Recurrent with attention [27]	84.9
Data augmentation [47]	88.7
Harmonic embeddings [26]	89.9
Synthesis data [48]	90.0
PFFNet	**91.1**

the previous state-of-the-art methods. However, the image synthesis methods in [47] and [48] are entirely based on the characteristic of the leaves in the given plant phenotype images, such as the texture, direction, and spatial relationship with other leaves. Therefore, the methods are task-specific and hard to fit to other datasets with different characteristics. With the help of panoptic-level features in a local and global view, our proposed-based PFFNet outperforms all the other methods on the CVPPP A1 dataset, without any task-specific design. In addition, the competitive performance on other instance segmentation tasks further demonstrates the generalization ability of our PFFNet.

E. Comparison With State-of-the-Art Panoptic Segmentation Methods for General Images

In Section IV-D, we compare our PFFNet with state-of-the-art methods, particularly on biomedical and biological image instance segmentation tasks. To further demonstrate the superiority of our proposed PFFNet, we conduct extensive experiments in comparison with state-of-the-art panoptic segmentation methods, including the original Panoptic Segmentation method (OGPan) [19], Panoptic FPN (PanFPN) [20], JSIS-Net [31], AUNet [32], UPSNet [49], and OANet [33]. Since these methods were originally proposed for general image analysis, we reimplement them on our tasks by following their original paper and source code. For fair comparisons, we maintain the same ResNet101 + FPN backbone and the group normalization strategy for these methods as ours. The inference processes of all the compared methods are also similar to ours, which employ the output of the instance branch as the final predictions. The experiment results are shown in Table VIII, where our proposed PFFNet outperforms all the compared methods on four instance segmentation tasks for biomedical and biological images.

Among the compared methods, only OGPan [19] proposes to train the semantic and instance segmentation tasks separately, which achieves the same performance as Mask R-CNN mentioned in Section IV-D. By jointly optimizing semantic and instance segmentation branches with a shared backbone, our PFFNet is able to induce the instance branch to process the semantic-level contextual information and therefore achieves better performance than the OGPan method. JSIS-Net [31] and Panoptic FPN [20] jointly train an instance branch with a semantic segmentation decoder, without any fusion mechanism between the features from the two branches. By fusing the features for each instance mask prediction to a semantic features map based on its corresponding location prediction, the instance branch in our PFFNet directly involved in the optimization process of the semantic segmentation tasks, which further facilitates the contextual feature learning and achieves better performance.

By integrating the information from the semantic and instance branch, AUNet [32], UPSNet [49], and OANet [33] achieve better performance than Panoptic FPN, JSIS-Net, and OGPan. AUNet [32] fuses the features from RPN and instance segmentation predictions with the different stages of the semantic branch via attention mechanism. Although the motivation of the attention feature fusion step in AUNet is similar to our residual attention feature fusion mechanism, the effectiveness of the semantic consistency regularization and the mask quality sub-branch still induce our PFFNet to achieve higher accuracies. For UPSNet [49], a panoptic segmentation head has been proposed to learn the panoptic-level information from the fused semantic- and instance-level features, which treats each instance as a unique category. However, the scale variation of the objects in the biomedical and biological images makes the size of each instance category vary. This brings the class-imbalance issue for optimizing the panoptic head of the UPSNet and degrades the overall performance. To this end, UPSNet achieves less competitive performance than our PFFNet due to its imperfect design under the biomedical and biological image instance segmentation application settings. In OANet [33], a spatial ranking module is proposed to predict a confidence score for each instance mask prediction during inference, which aims at tackling the objects overlapping issue when small instances stay inside larger ones. However, such a situation might rarely exist in biomedical and biological images, as shown in Fig. 5, 7, 8, and 9. Subsequently, the performance gain of the spatial ranking module is limited to our tasks. On the other hand, our mask quality sub-branch is based on the mask segmentation quality for each instance, which avoids the influence from the overlapping situations and maintains the competitive performance of our PFFNet.

F. Ablation Study

In this section, we conduct ablation experiments on the TNBC, BBBC039V1, and CVPPP dataset, to test the effectiveness of the three newly proposed modules in the PFFNet on different types of images. For TNBC and BBBC039V1, we have the same data settings as the previous experiments. For CVPPP, we conduct 3-fold cross-validation on the 128 training images.

1) Residual Attention Feature Fusion Mechanism: In this section, we first study the selection of the feature fusion mechanism. As shown in Fig 10, we present 4 different selections: (a) replacement: replace the semantic feature map with the mask predictions, which is employed in Cell R-CNN V2; (b) summation: sum the semantic feature map with the mask prediction; (c) attention: we first obtain the mask probability

TABLE VIII

THE COMPARISON EXPERIMENTS BETWEEN OUR PROPOSED PFFNET AND THE STATE-OF-THE-ART PANOPTIC SEGMENTATION METHODS FOR GENERAL IMAGES

	Kumar				*TNBC*				*BBBC039V1*				*CVPPP*		
	AJI	*Dice*	*F1*	*PQ*	*AJI*	*Dice*	*F1*	*PQ*	*AJI*	*Dice*	*F1*	*PQ*	*SBD*	*Dice*	*PQ*
OGPan [19]	0.5396	0.7659	0.6750	0.4796	0.5350	0.7393	0.7542	0.5146	0.7983	0.9277	0.9180	0.7773	0.8674	0.8453	0.6342
JSISNet [31]	0.5653	0.7664	0.7902	0.5448	0.5673	0.7601	0.7776	0.5432	0.8134	0.9316	0.9282	0.7913	0.8835	0.8799	0.6848
PanFPN [20]	0.5811	0.7898	0.8194	0.5678	0.5874	0.7715	0.8129	0.5720	0.8193	0.9320	0.9275	0.7960	0.8871	0.8905	0.6999
AUNet [32]	0.5898	0.7904	0.8056	0.5650	0.5932	0.7761	0.8147	0.5793	0.8252	0.9377	0.9315	0.8090	0.8883	0.9086	0.7153
UPSNet [49]	0.5797	0.7904	0.8315	0.5667	0.5816	0.7703	0.8125	0.5625	0.8128	0.9274	0.9191	0.7857	0.8902	0.8941	0.7180
OANet [33]	0.5865	0.7908	0.8167	0.5645	0.5933	0.7798	0.8126	0.5744	0.8198	0.9372	0.9330	0.8085	0.8881	0.8994	0.7102
PFFNet	**0.6107**	**0.8091**	**0.8345**	**0.5871**	**0.6313**	**0.8037**	**0.8600**	**0.6298**	**0.8477**	**0.9478**	**0.9451**	**0.8330**	**0.9062**	**0.9338**	**0.7788**

TABLE IX

THE COMPARISON OF RESULTS FOR THE DIFFERENT DESIGN SELECTIONS FOR THE FEATURE FUSION BETWEEN THE SEMANTIC AND INSTANCE FEATURES. THE (A), (B), (C), AND (D) CORRESPOND TO FIG. 10

	TNBC				*BBBC039V1*				*CVPPP*		
	AJI	*Dice*	*F1*	*PQ*	*AJI*	*Dice*	*F1*	*PQ*	*SBD*	*Dice*	*PQ*
(a)	0.6133	0.791	0.8321	0.5984	0.8335	0.9410	0.9407	0.8205	0.8947	0.9200	0.7491
(b)	0.6216	0.800	0.8541	0.6179	0.8420	0.9443	0.9407	0.8276	0.8983	0.9274	0.7603
(c)	0.6220	0.7983	0.8521	0.6139	0.8403	0.9443	0.9427	0.8248	0.8999	0.9267	0.7606
(d)	**0.6313**	**0.8037**	**0.8600**	**0.6298**	**0.8477**	**0.9478**	**0.9451**	**0.8330**	**0.9062**	**0.9338**	**0.7788**

TABLE X

THE ABLATION STUDY FOR THE MASK QUALITY SUB-BRANCH. W/O MEANS PFFNET WITHOUT THE MASK QUALITY SUB-BRANCH

	TNBC				*BBBC039V1*				*CVPPP*		
	AJI	*Dice*	*F1*	*PQ*	*AJI*	*Dice*	*F1*	*PQ*	*SBD*	*Dice*	*PQ*
w / o	0.6150	0.7960	0.8495	0.6114	0.8387	0.9426	0.9388	0.8206	0.8977	0.9233	0.7593
w	**0.6313**	**0.8037**	**0.8600**	**0.6298**	**0.8477**	**0.9478**	**0.9451**	**0.8330**	**0.9062**	**0.9338**	**0.7788**

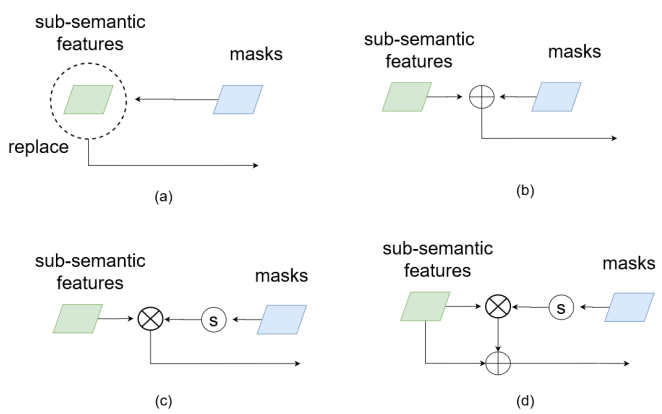

Fig. 10. Various fusion methods of the proposed residual attention feature fusion mechanism. S represents the sigmoid operation. $+$ is the element-wise summation and $\times$ is the element-wise multiplication.

maps with the sigmoid operation, then they are multiplied with the corresponding semantic features; (d) residual attention: add a residual connection on (c). For all the experiments, the rest of the model is the same as our proposed PFFNet in addition to the feature fusion mechanism. The comparison results are shown in Table IX.

As discussed above, the previous feature fusion mechanism in Cell R-CNN V2 directly replaced the subset of the semantic feature map using the mask predictions for each object, which results in the semantic-level information loss. Therefore, the models with summation (b), attention (c), and residual attention (d) fusion mechanism achieve better performance than the replacement fusion (a), as all of them preserve the original semantic features. Among (b), (c), and (d), residual attention fusion mechanism (d) always achieves the best performance on all datasets, and it is therefore employed in our PFFNet.

2) Mask Quality Sub-Branch: To demonstrate the effectiveness of our proposed mask quality sub-branch, we conducted an ablation study by removing the mask quality sub-branch and comparing the performance with PFFNet. As shown in Table X, the accuracies under all the metrics are decreased after removing the mask quality sub-branch, especially on the object-level metrics. Without the mask quality sub-branch, there exist low-quality mask predictions with a high classification score, which affect the segmentation of the touching object during inferences and are eventually harmful to the object-level performance.

3) Semantic Task Consistency Regularization: Table XI illustrates the effectiveness of the semantic consistency regularization by ablating it from the original PFFNet. Although the regularization aims to facilitate the semantic information learning in the instance branch, we notice that the improvements under the object-level metrics are at the same level of the pixel-level metrics, on all the three datasets. Compared with Table IX and Table X, we notice that the model without the semantic consistency regularization outperforms the model without the residual attention fusion mechanism and mask quality sub-branch, which indicates the effectiveness of the semantic consistency regularization is the lowest among all

TABLE XI

THE ABLATION STUDY FOR THE SEMANTIC TASK CONSISTENCY REGULARIZATION. W/O MEANS PFFNET WITHOUT THE SEMANTIC CONSISTENCY REGULARIZATION

	TNBC				*BBBC039V1*				*CVPPP*		
	AJI	*Dice*	*F1*	*PQ*	*AJI*	*Dice*	*F1*	*PQ*	*SBD*	*Dice*	*PQ*
w / o	0.6206	0.7974	0.8370	0.6043	0.8390	0.9430	0.9416	0.8244	0.8995	0.9267	0.7601
w	**0.6313**	**0.8037**	**0.8600**	**0.6298**	**0.8477**	**0.9478**	**0.9451**	**0.8330**	**0.9062**	**0.9338**	**0.7788**

TABLE XII

THE EXPERIMENTAL RESULTS ON THE MODEL GENERALIZATION ABILITY STUDY FOR OUR PROPOSED PFFNET

Method	*AJI*	*Dice*	*SQ*	*DQ*	*PQ*
Cell R-CNN V2 [22]	0.5838	0.8006	0.7495	0.7601	0.5713
PFFNet	0.6048	0.8146	0.7559	0.7729	0.5856
NuClick [50]	0.7940	0.8886	0.8001	0.9819	0.7856

the three proposed modules. However, the semantic task consistency regularization is still a novel module as it is implemented by only adding one more loss function, which is straightforward and easy to adapt to other related tasks.

G. Generalization Ability Study

To demonstrate the generalization ability of our proposed PFFNet, we conduct a generalization study by training the model on one seen dataset and validate it on another unseen one. We follow the experimental setting in [50] by training the model on the CPM17 training set [51] and validating it on the TCGA-Kumar testing set. During training, we first randomly cropped 10 256 $\times$ 256 patches from each CPM17 training images. Next, the patches are augmented via horizontal and vertical flipping, 90°, 180°, and 270° rotations, Gaussian noise, Gaussian blur, and median blur. For testing, the well-trained model is validated on the 14 testing images from the TCGA-Kumar testing set. The detailed experimental results are illustrated in Table XII.

As shown in Table XII, our proposed PFFNet achieves better performance than our previous Cell R-CNN V2 [22] under all metrics. By enhancing the generalization ability of the Cell R-CNN V2, our proposed residual attention feature fusion mechanism, mask quality branch, and semantic task consistency regularization module are further demonstrated to be effective. In addition, we also compare with Nuclick [50], which employs the point annotations of the nuclei instance masks during training and inference. Without accessing the point annotations for the testing images, our PFFNet achieves less competitive performance than Nuclick. However, we notice that the performance of our PFFNet is close to NuClick under Dice score for pixel-level segmentation, and the SQ score for instance-level segmentation. On the other hand, acquiring the nuclei point annotations for the unseen datasets still incurs annotation burdens. If a less competitive accuracy is acceptable, we would prefer not to employ any annotations from the unseen testing datasets, in favour of a low manual annotation cost.

V. CONCLUSION

In this work, we propose a novel Panoptic Feature Fusion Net (PFFNet) for instance segmentation in the biomedical and biological images, which incorporates semantic- and instance-level features. By extending our previous Cell R-CNN [21] and Cell R-CNN V2 [22], our newly proposed PFFNet is improved with a residual attention feature fusion mechanism, mask quality sub-branch, and semantic consistency regularization. With the help of the residual attention feature fusion mechanism, the semantic features of foreground objects are retained and the mask generators are able to learn more global contextual information from the semantic segmentation task in the instance branch. To alleviate the misalignment issue between the quality of each mask prediction and its classification score, our mask quality sub-branch learns the mask prediction quality scores during training and employs the quality score to re-weigh the classification of each instance prediction during inference. For robust and accurate learning of the semantic features, the semantic consistency mechanism is proposed to regularize the two semantic segmentation tasks jointly. Furthermore, our PFFNet has wide applicability on various biomedical and biological datasets, including histopathology images, fluorescence microscopy images, and plant phenotype images, where we outperform several state-of-the-art methods by a large margin, including our previous Cell R-CNN and Cell R-CNN V2.

By fulfilling the future work in Cell R-CNN V2 [22], our PFFNet has been verified to be effective on various biomedical and biological datasets. In future work, we would further adapt our PFFNet to general image processing tasks. As our PFFNet is effective for 2D image analysis, we can also extend it for 3D microscopy image instance segmentation, which is another important and interesting problem related to this work.

REFERENCES

[1] C. W. Elston and I. O. Ellis, "Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: Experience from a large study with long-term follow-up," *Histopathology*, vol. 19, no. 5, pp. 403–410, Nov. 1991.

[2] F. Clayton, "Pathologic correlates of survival in 378 lymph node-negative infiltrating ductal breast carcinomas. Mitotic count is the best single predictor," *Cancer*, vol. 68, no. 6, pp. 1309–1317, Sep. 1991.

[3] A. Madabhushi *et al.*, "Multi-field-of-view strategy for image-based outcome prediction of multi-parametric estrogen receptor-positive breast cancer histopathology: Comparison to oncotype DX," *J. Pathol. Informat.*, vol. 2, no. 2, p. 1, 2011.

[4] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: A comprehensive review," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 234–263, 2016.

[5] M. Minervini, A. Fischbach, H. Scharr, and S. A. Tsaftaris, "Finely-grained annotated datasets for image-based plant phenotyping," *Pattern Recognit. Lett.*, vol. 81, pp. 80–89, Oct. 2016.

[6] H. Scharr *et al.*, "Leaf segmentation in plant phenotyping: A collation study," *Mach. Vis. Appl.*, vol. 27, no. 4, pp. 585–606, May 2016.

[7] H. Scharr, M. Minervini, A. Fischbach, and S. A. Tsaftaris, "Annotated image datasets of rosette plants," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 6–12.

[8] H. Llewellyn, "Observer variation, dysplasia grading, and HPV typing," *Pathol. Patterns Rev.*, vol. 114, pp. S21–S35, Nov. 2000.

[9] J. Song, L. Xiao, and Z. Lian, "Contour-seed pairs learning-based framework for simultaneously detecting and segmenting various overlapping cells/nuclei in microscopy images," *IEEE Trans. Image Process.*, vol. 27, no. 12, pp. 5759–5774, Dec. 2018.

[10] C. Payer, D. Štern, M. Feiner, H. Bischof, and M. Urschler, "Segmenting and tracking cell instances with cosine embeddings and recurrent hourglass networks," *Med. Image Anal.*, vol. 57, pp. 106–119, Oct. 2019.

[11] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Trans. Med. Imag.*, vol. 36, no. 7, pp. 1550–1560, Jul. 2017.

[12] P. Naylor, M. Lae, F. Reyal, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE Trans. Med. Imag.*, vol. 38, no. 2, pp. 448–459, Feb. 2019.

[13] C. Payer, D. Štern, T. Neff, H. Bischof, and M. Urschler, "Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[14] B. De Brabandere, D. Neven, and L. Van Gool, "Semantic instance segmentation with a discriminative loss function," 2017, *arXiv:1708.02551*. [Online]. Available: http://arxiv.org/abs/1708.02551

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

[16] D. Zhang, Y. Song, S. Liu, D. Feng, Y. Wang, and W. Cai, "Nuclei instance segmentation with dual contour-enhanced adversarial network," in *Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2018, pp. 409–412.

[17] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2017, pp. 2980–2988.

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

[19] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9404–9413.

[20] A. Kirillov, R. Girshick, K. He, and P. Dollar, "Panoptic feature pyramid networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6399–6408.

[21] D. Zhang *et al.*, "Panoptic segmentation with an end-to-end cell R-CNN for pathology image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2018, pp. 237–244.

[22] D. Liu *et al.*, "Nuclei segmentation via a deep panoptic model with semantic feature fusion," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, AAAI Press, Aug. 2019, pp. 861–868.

[23] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring R-CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6409–6418.

[24] M. Bai and R. Urtasun, "Deep watershed transform for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5221–5229.

[25] V. Kulikov, V. Yurchenko, and V. Lempitsky, "Instance segmentation by deep coloring," 2018, *arXiv:1807.10007*. [Online]. Available: http://arxiv.org/abs/1807.10007

[26] V. Kulikov and V. Lempitsky, "Instance segmentation of biological images using harmonic embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3843–3851.

[27] M. Ren and R. S. Zemel, "End-to-end instance segmentation with recurrent attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6656–6664.

[28] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "MaskLab: Instance segmentation by refining object detection with semantic and direction features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4013–4022.

[29] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 28, 2019, doi: 10.1109/TPAMI.2019.2956516.

[30] F. S. Saleh, M. S. Aliakbarian, M. Salzmann, L. Petersson, and J. M. Alvarez, "Effective use of synthetic data for urban scene semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 86–103.

[31] D. de Geus, P. Meletis, and G. Dubbelman, "Panoptic segmentation with a joint semantic and instance segmentation network," 2018, *arXiv:1809.02110*. [Online]. Available: http://arxiv.org/abs/1809.02110

[32] Y. Li *et al.*, "Attention-guided unified network for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7026–7035.

[33] H. Liu *et al.*, "An end-to-end network for panoptic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6172–6181.

[34] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4974–4983.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[36] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1743–1751.

[37] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[38] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[39] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool, "Domain adaptive faster R-CNN for object detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3339–3348.

[40] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter, "Annotated high-throughput microscopy image sets for validation," *Nat. Methods*, vol. 9, no. 7, p. 637, 2012.

[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[43] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 3–19.

[44] A. Paszke *et al.*, "Automatic differentiation in pytorch," in *Proc. Conf. Neural Inf. Process. Syst. Autodiff Workshop*, 2017. [Online]. Available: https://autodiff-workshop.github.io/ and https://openreview.net/forum?id=BJJsrmfCZ

[45] B. Romera-Paredes and P. H. S. Torr, "Recurrent instance segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 312–329.

[46] A. Salvador *et al.*, "Recurrent neural networks for semantic instance segmentation," 2017, *arXiv:1712.00617*. [Online]. Available: http://arxiv.org/abs/1712.00617

[47] D. Kuznichov, A. Zvirin, Y. Honen, and R. Kimmel, "Data augmentation for leaf segmentation and counting tasks in rosette plants," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[48] D. Ward, P. Moghadam, and N. Hudson, "Deep leaf segmentation using synthetic data," in *Proc. BMVC*, 2018, pp. 327–339. [Online]. Available: http://bmvc2018.org/contents/workshops/cvppp2018/0026.pdf

[49] Y. Xiong *et al.*, "UPSNet: A unified panoptic segmentation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8818–8826.

[50] M. Jahanifar, N. Alemi Koohbanani, and N. Rajpoot, "NuClick: From clicks in the nuclei to nuclear boundaries," 2019, *arXiv:1909.03253*. [Online]. Available: http://arxiv.org/abs/1909.03253

[51] Q. D. Vu *et al.*, "Methods for segmentation and classification of digital microscopy tissue images," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 53, Apr. 2019.