Hindawi
Journal of Sensors
Volume 2021, Article ID 8872947, 8 pages
https://doi.org/10.1155/2021/8872947

Research Article

Weighted Mask R-CNN for Improving Adjacent Boundary Segmentation

SungMin Suh,[1] **Yongeun Park**,[2] **KyoungMin Ko**,[1] **SeongMin Yang**,[1,3] **Jaehyeong Ahn**,[1] **Jae-Ki Shin**,[4] **and SungHwan Kim**[1,3]

[1]*Department of Applied Statistics, Konkuk University, Seoul, Republic of Korea*
[2]*Department of Social and Environmental Engineering, Konkuk University, Seoul, Republic of Korea*
[3]*AI Analytics Team, DeepVisions, Seoul, Republic of Korea*
[4]*Korea Water Resources Corporation, Busan, Republic of Korea*

Correspondence should be addressed to Jae-Ki Shin; shinjaeki@gmail.com and SungHwan Kim; shkim1213@konkuk.ac.kr

Received 16 August 2020; Revised 23 November 2020; Accepted 14 December 2020; Published 23 January 2021

Academic Editor: Zhenxing Zhang

In the recent era of AI, instance segmentation has significantly advanced boundary and object detection especially in diverse fields (e.g., biological and environmental research). Despite its progress, edge detection amid adjacent objects (e.g., organism cells) still remains intractable. This is because homogeneous and heterogeneous objects are prone to being mingled in a single image. To cope with this challenge, we propose the weighted Mask R-CNN designed to effectively separate overlapped objects in virtue of extra weights to adjacent boundaries. For numerical study, a range of experiments are performed with applications to simulated data and real data (e.g., *Microcystis*, one of the most common algae genera and cell membrane images). It is noticeable that the weighted Mask R-CNN outperforms the standard Mask R-CNN, given that the analytic experiments show on average 92.5% of precision and 96.4% of recall in algae data and 94.5% of precision and 98.6% of recall in cell membrane data. Consequently, we found that a majority of sample boundaries in real and simulated data are precisely segmented in the midst of object mixtures.

1. Introduction

The identification of genera in water samples is of central importance in assessing water quality in vision. Over the years, this procedure has mainly relied on manual counting [1], which inevitably suffers limitations in consuming time, manpower, and energy. Thus, it is urgent to develop vision sensing-based automatic tools capable of expediting the detection and quantification process. Commonly, previous studies on algae genera have focused on developing accurate classification models. For identifying labels, the model is designed to predict the corresponding taxa, learning on images containing genera of interest. Large-scale data by augmentation technique have been exploited to fine-tune a model on the basis of the AlexNet architecture [2]. It is remarkable that they have achieved performance of overall accuracy 99.51% of 80 genera, each of which contains more than 2000 samples. Different from deep learning-based methods, various predictive models based on hand-crafted features also reported promising results. Importantly, Schulze et al. and Bueno et al. [3, 4] have obtained 95% and 98% accuracy, respectively. Given that the accuracies of the studies nearly come to 100%, seemingly it seems that classification of genera is conquered. Apart from this, Park et al. [5] have proposed the Bayesian optimization-based neural architecture search (BO-NAS) for a better classification of cyanobacteria with the convolutional neural networks (CNN). Using the flow cytometer and microscope (FlowCAM; [6]), they collected the image data of cyanobacteria including *Microcystis* characterized in interfering effects due to crowded cells and diatoms. It is remarkable that this CNN model effectively classified the algal genus with an F1 score,

which is the harmonic mean of precision and recall, of 0.95 for the eight genera. Interestingly, leveraging all of the CNN, the grayscale surface direction angle model (GSDAM; [7]) and Canny edge detection [8, 9] have identified algae in an unsupervised fashion. Mary and Prabakaran [10] segmented and classified 70 genera of 1531 images using Canny edge detection and the Inception V4 [11]. Previous studies have achieved significant classification results on some genera images, but they were limited in scope to classification [12]. To detect and quantify genera furthermore, several intractable problems still remained. As discussed in [1], it is required to locate genera presented in the image since the taxonomist handles images containing multiple taxa. For doing this, we necessarily introduce both Region of Interest (ROI) detection and instance segmentation algorithm.

Recently, image classification has been applied in a variety of fields such as geoscience and remote sensing (RS). In the hyperspectral (HS) images containing specialties on spatial information, several research projects have been successfully made [13]. Hong et al. [14] address the HS images focusing on RS images with the multimodal deep learning framework (MDL-RS). The MDL-RS networks propose five plug-and-play fusion modules making possible to submit the image information effectively through the modalities. In the two extraction subnetworks (Ex-Net) based on pixel-wise or spatial-spectral architectures, each modality extracts the feature map through the CNN-based networks. Embedding the Ex-Net outputs to the input of fusion network (Fu-Net), the Fu-Net binds the feature maps using concatenation- and compactness-based methods. The nonlocal graph convolutional network (nonlocal GCN) classifies the HS images with a novel graph-based semisupervised learning [15].

Furthermore, the recent studies also pay attention to detecting the precise boundary in the midst of the complex image data. Xie et al. [16] utilize the hyperparameters to train and used transfer learning to reduce the training time of the GlacierNet CNN modified from the SegNet [17]. In [18], the deep fully convolutional network dilated kernel (FCN-DK) based on the supervised pixel-wise image classification for improving cadastral boundary detection in urban and semi-urban areas is proposed. The performance of the model is compared with the state-of-the-art techniques, including Multiresolution Segmentation (MRS; [19]) and Globalized Probability of Boundary (gPb; [20]). For the medical image segmentation especially in CT images, the adaptive fully dense (AFD) neural network adding the horizontal connections in U-Net structure [21] is known to perform outstanding boundary detection [22].

Instance segmentation is the simultaneous task of detecting and delineating each distinguishable object in an image. Breaking through the Faster R-CNN [23], the model used object detection with a parallel branch for predicting segmentation masks, namely, the Mask R-CNN [24], surpassing all the previous state-of-the-art methods on the COCO instance segmentation data set [25], and has been widely applied to diverse academic domain. Although its superior performance is unquestionable, it still has difficulty in handling densely crowded and overlapping instances.

To address these obstacles, we propose a novel way of improving the Mask R-CNN by accommodating extra weights in the model that integrates prior known knowledge. In the experiments, we apply weights to neighboring boundaries of algae especially in *Microcystis* genus which are quite complex to classify because of the variety form of algae. Notably, it is also shown worthy of effectively counting cells (i.e., vision sensing) through calculating objective areas for the measurement of concentration in algae. Moreover, we leverage heavy weights to adjacent boundaries of objects in multiple cell membrane images for improved accuracy.

The rest of this paper is organized as follows. In Section 2, the proposed methods are given. Next, in Section 3, we describe how we acquire the image data sets, preprocess, and provide experiment results. In Section 4, we discuss our results comparing with existing works and address future works.

2. Methods

2.1. Mask R-CNN Network Architecture. Network architectures of the Mask R-CNN largely consist of two parts: (1) feature extraction and (2) instance segmentation. First, the ResNet101 module [24] pretrained by the COCO data set is used. The backbone network and feature pyramid network (FPN) architecture designed to extract features are used for better accuracy and processing speed. Next, in the head of network architecture, the model detects ROI, and from the derived ROI detection and classification are made. With these frameworks, the fully convolutional mask prediction is lastly implemented for instance segmentation.

2.2. Integration of Distance Weight with Mask R-CNN. Here, standing on the shoulder of the Mask R-CNN, we propose the weighted Mask R-CNN specially designed to accommodate *a priori* known weights to the main objective function. This method is mainly aimed at precisely separating the boundary of multiple samples in the context of instance segmentation. Putting in a nutshell, the tasks of the Mask R-CNN achieve largely three goals: (1) classifying class labels, (2) detecting bounding boxes, and (3) segmenting instances. Firstly, the model extracts feature maps by passing resized images through the CNN. On the basis of the feature maps, the Region Proposal Network (RPN) stage allows for the candidates of objective bounding box among generated anchor boxes. Subsequent to this, the ROI align is performed to gather the precise pixel location data. The ROI align serves as a building block to detect objects as well as to segment instances. Focusing on ROI align, the model extracts feature maps of interest areas by using exact coordinates through fully convolutional network (FCN [26]). Afterwards, through the process of minimizing the objective function, we optimize the Mask R-CNN model. The model defines the objective function as the aggregation of the loss functions of classification, localization, and segmentation [27]. Moreover, each loss function is optimized by the softmax function, box offset regressor, and mask FCN predictor, respectively. In this process, the novelty of the Mask R-CNN comes into play in advancing the former image recognition models (e.g., Fast

R-CNN [28] and Faster R-CNN). While deriving the objective function, the Mask R-CNN implements the pixel-wise binary classification and decouples mask prediction with both category classification and bounding box detection. Notably, the binary classification method has merits in terms of reduction computation costs. The ROI align precisely masks, aiming at approximating ground truth areas.

For the weighted Mask R-CNN, below is the proposed objective function:

$$\begin{aligned} L &= L_{\text{cls}} + L_{\text{box}} + w \cdot L_{\text{mask}} \\ &= \frac{1}{N_{\text{cls}}} \sum_i \{-p_i^* \cdot \log p_i - (1 - p_i^*) \cdot \log (1 - p_i)\} + \frac{\lambda}{N_{\text{box}}} \sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*) \\ &\quad - w \cdot \frac{1}{m^2} \sum_{1 \le i,j \le m} \left\{ y_{ij} \cdot \log \hat{y}_{ij}^k + \left(1 - y_{ij}\right) \cdot \log \left(1 - \hat{y}_{ij}^k\right) \right\}, \end{aligned} \tag{1}$$

where p_i is the predicted probability of anchor i being an object, p_i^* is the ground truth label (binary) of whether anchor i is an object, t_i is the predicted four parameterized coordinates, t_i^* is the ground truth coordinates, N_{cls} is the normalization term set to be minibatch size (0~256), N_{box} is the normalization term set to the number of anchor locations (0~2400), λ is the balancing parameter set to be (0~10 such that both L_{cls} and L_{box} terms are roughly, equally weighted), k is the number of ground truth class, w is the weight matrix assigned to pixel instances, and

$$L_1^{\text{smooth}}(x) = \begin{cases} 0.5x^2, & \text{if } |x| \le 1, \\ |x| - 0.5, & \text{otherwise.} \end{cases} \tag{2}$$

In addition, we integrate both image representations and *a priori* known knowledge of adjacency in the model. Inspired by the U-Net, this weight induces strong separation across samples as boundaries get closer. In theory, the closer the boundary the bigger the weight:

$$\begin{aligned} w(x) &= w_c(x) + w_0 \cdot e^{-(d_1(x) + d_2(x))^2 / 2\sigma^2}, \\ w &= 1 + \frac{w(x) - \min\left(w(x)\right)}{\max\left(w(x)\right) - \min\left(w(x)\right)} \cdot \delta, \end{aligned} \tag{3}$$

where $w_c : \Omega \to R$ is the weight map to balance the class frequencies, $d_1 : \Omega \to R$ denotes the distance to the border of the nearest cell, $d_2 : \Omega \to R$ denotes the distance to the border of the second nearest cell, and δ refers to the weight adjusting parameter, respectively.

In principle, $w(x)$ is subject to size of objects, distance between objects, and shape of the objects in an image. To account for variability, we scale each weight map separately to the range from 0 to 1. Next, we consider the parameter δ to determine the power of the weight matrix. The weight parameter δ can be used for adding the extra emphasis on the boundary of object especially when the distance between objects is too narrow so that we hardly distinguish boundaries. Subsequent to this, we impose this weight matrix to the objective function of masks in the fashion of element-wise computation. Taken together, Figure 1 displays the end-to-end architecture of the proposed model.

Moreover, the stochastic gradient descent (SGD) algorithm is used as an optimizer and minibatch size is fixed to 1 in this study, and we set the learning rate of 0.001 and 100 epochs. Validation processes with comparing ground truth masks to assess predictive performance. For implementation, the Mask R-CNN adopts the PyTorch packages for simplicity [29].

3. Numerical Experiments

3.1. Data Sets. In what follows, we describe the data sets for numerical study. First and foremost, it is essential to generate well-preprocessed data sets to produce reliable experiment results. To this end, we apply several preprocessing techniques such as standardization or scaling to raw data and matching each preprocessed image with precise annotations.

3.1.1. Simulated Data. In simulation I, we generate circle images each of which includes inside 4 and 6 circled objects for train data sets, respectively, where all images have resolution of $512 \times 512 \times 3$ pixels. Similarly, we generate circle images including the prespecified number of objects for test data set (i.e., 4 and 6). Subsequent to this, we divide each image both in horizontal and vertical direction in the way that each circle is exclusively placed one at a diagonal slot and the radius of each circle is limited to the boundary of slots. Simulation II emulates the nature of real data, for which we generate the shape of ellipses in accommodating randomness and complexity to the simulation data sets. More precisely, we randomly choose the center points of objects and generate ellipses of random sizes for experiment data sets assigned to the diagonal slots. This configuration makes distance between objects arbitrarily determined and promotes adequate complexity.

3.1.2. Microalgae and Cell Membrane Data. Freshwater microalgae samples used in this work were collected at 11 weirpools and five reservoirs located in the four major rivers (e.g., Han, Nakdong, Geum, and Yeongsan) in Korea. Water (quantitative) or net (qualitative) samples were taken from the surface and immediately fixed to the final 1% concentration with acidified Lugol's iodine solution [30]. Quantitative samples were allowed to stand in the dark place of the laboratory for more than one week, and then, the supernatant was carefully siphoned and concentrated an appropriate cell density (above 104 cells/mL). Image acquisition was performed using photomicroscopes (Zeiss AXIO Scope.A1 and Vert.A1 model, Germany) attached camera (Axiocam 506 color) assisted with computer software (ZEN lite 2012), and captured images have resolution of $1936 \times 1460 \times 3$ pixels at 200x or 400x magnification of a microscope. A manual identification of algae species was carried out based on their taxonomic characteristics by [31].

In the experiment, 469 *Microcystis* images are used in total. Since the images are collected insufficiently, the performance of segmentation model can be severely deteriorated.

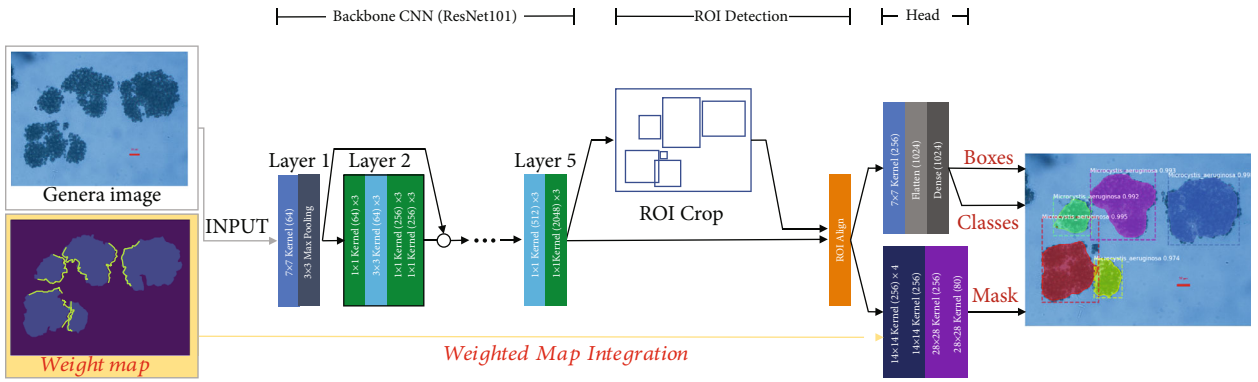

Figure 1: Integration of the weight map under the architecture of the Mask R-CNN.

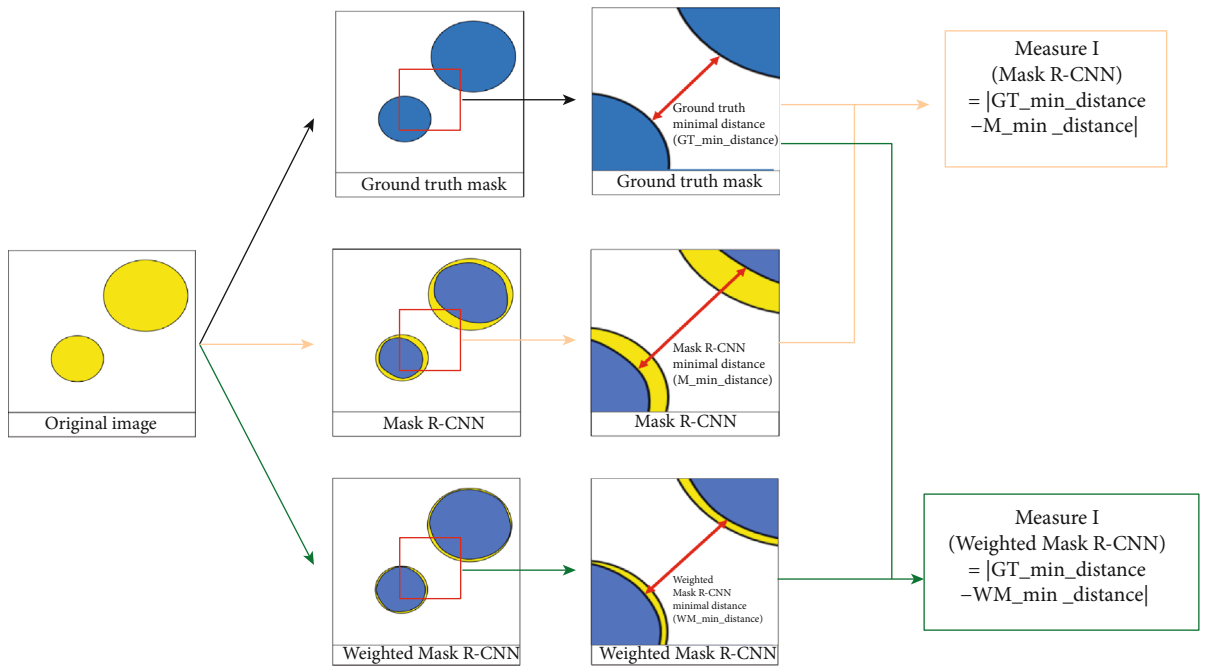

Figure 2: Examples of calculating Measure I from two ellipse images.

However, we fine-tune by means of the CNN pretrained with the COCO data set in order to tackle the degrading performance problem. In addition to this, we also analyze 30 cell membrane images in electron microscopic (EM; [32]) segmentation challenge at the International Symposium on Biomedical Imaging (ISBI). After that, taxonomists elaborately assess the consistency of labeling and annotations. LabelMe (https://github.com/wkentaro/labelme) is used as an annotation tool widely accepted for segmentation tasks. Importantly, it is very useful to annotate polygons simply by marking points and labeling genus taxa challenging due to the complexity and variety of shapes of algae and cell membrane. Thus, we annotate one by one to accurately delineate sophisticated boundaries. Finally, the annotation files are automatically saved in the JSON file format. For algae data set, we split the whole data into the training set of 319 images and the test set of 150 images.

3.2. Results

3.2.1. Evaluation Metrics. True positive (TP) pixels are ground truth target pixels and also predicted as target pixels. True negative (TN) pixels are not ground truth target pixels and also not predicted as target pixels. False positive (FP) pixels are not ground truth target pixels but predicted as target pixels known as Type II Error. False negative (FN) pixels are ground truth target pixels but not predicted as target pixels called as Type I Error. Precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

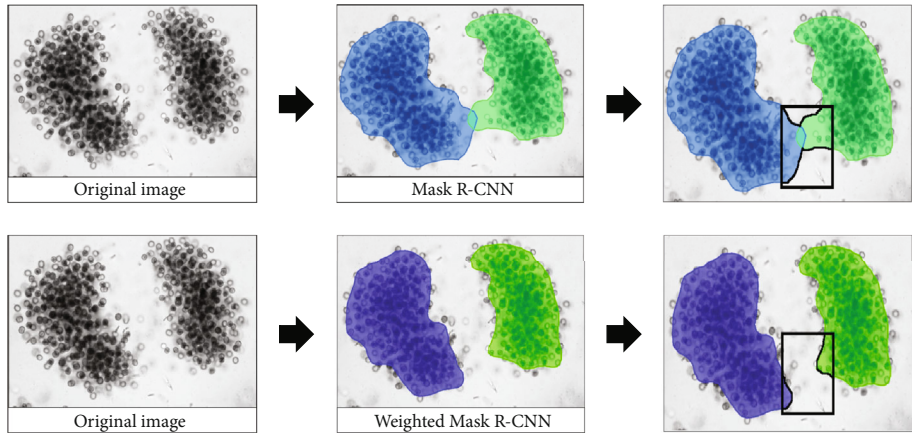

Figure 3: Examples of calculating Measure II from algae images.

Table 1: The comparison of Mask R-CNN, weighted Mask R-CNN, and MEInst in simulation I.

# of circles	4					6				
Model	MEInst	Mask R-CNN	Weighted Mask R-CNN			MEInst	Mask R-CNN	Weighted Mask R-CNN		
			$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$			$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$
Mean	0.768	1.110	0.629	0.857	0.984	0.834	0.936	0.659	0.830	1.008
se	0.062	0.056	0.048	0.052	0.052	0.036	0.034	0.031	0.035	0.033

Table 2: The comparison of Mask R-CNN, weighted Mask R-CNN, and MEInst in simulation II.

# of ellipses	4					6				
Model	MEInst	Mask R-CNN	Weighted Mask R-CNN			MEInst	Mask R-CNN	Weighted Mask R-CNN		
			$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$			$\delta = 0.1$	$\delta = 0.5$	$\delta = 1$
Mean	1.212	1.154	0.787	0.941	1.039	0.895	0.947	0.745	0.941	0.942
se	0.073	0.068	0.055	0.047	0.062	0.068	0.075	0.068	0.075	0.059

Table 3: Comparison of the performance of object detection between Mask R-CNN and weighted Mask R-CNN in algae and cell membrane images.

	Mask R-CNN		Weighted Mask R-CNN	
	Algae	Cell membrane	Algae	Cell membrane
mAP50	0.862	0.899	0.925	0.945
mAP75	0.688	0.796	0.786	0.852
mAP	0.603	0.625	0.673	0.677
Recall50	0.945	0.970	0.964	0.986
Recall75	0.785	0.867	0.843	0.908
Recall	0.699	0.704	0.734	0.739
mIoU	0.801	0.579	0.845	0.636

A precision-recall curve is a plot of precision (y-axis) and recall (x-axis) with varied thresholds. Average Precision (AP) is the under area of the precision-recall curve and is calculated as the mean precision given recall measures. mAP is the mean of Average Precision calculated by the multiple objects in an image. Intersection over Union (IoU) is a well-known measure from ground truth mask (gt_{mask}) and predicted mask ($\text{pred}_{\text{mask}}$) in evaluating image segmentation methods:

$$\text{IoU} = \frac{\text{gt}_{\text{mask}} \cap \text{pred}_{\text{mask}}}{\text{gt}_{\text{mask}} \cup \text{pred}_{\text{mask}}}. \quad (5)$$

In this study, we further define the mean IoU of multiple objects in an image (mIoU (mean of Intersection over Union)). In this paper, we compute mAP and recall at the given IoU threshold (default 0.5). Without the given IoU threshold, we compute mAP and recall over a range of IoU threshold (as default 0.5 to 0.95 with an increment of 0.05).

The first measurement in the boundary detection in this paper called as Measure I is defined as the absolute value of difference, that is, a minimal distance of adjacent two objects between the ground truth mask and predicted mask. Figure 2 describes the example of Measure I.

The second measurement in the boundary detection called as Measure II gauges the proportion of mask pixels among predesignated areas. We compare Measure II in both models with algae and cell membrane images, where the Mask R-CNN produces overlapping inferred masks of two

Table 4: Comparison between Mask R-CNN and weighted Mask R-CNN by Measure II in algae and cell membrane images.

Image names	Mask R-CNN	Weighted Mask R-CNN	Image names	Mask R-CNN	Weighted Mask R-CNN
Microcystis 1	0.38	0.09	Cell membrane 1	0.81	0.68
Microcystis 2	0.58	0.23	Cell membrane 2	0.82	0.82
Microcystis 3	0.82	0.74	Cell membrane 3	0.81	0.75
Microcystis 4	0.85	0.45	Cell membrane 4	0.68	0.60
Microcystis 5	0.41	0.18	Cell membrane 5	0.91	0.73
Microcystis 6	0.61	0.30	Cell membrane 6	0.62	0.50
Microcystis 7	0.77	0.35	Cell membrane 7	0.65	0.46
Microcystis 8	0.66	0.36	Cell membrane 8	0.80	0.74
Microcystis 9	0.80	0.18	Cell membrane 9	0.76	0.52
Microcystis 10	0.72	0.72	Cell membrane 10	0.86	0.75
Microcystis 11	0.57	0.31	Cell membrane 11	0.88	0.76
Microcystis 12	0.59	0.35	Cell membrane 12	0.35	0.34
Microcystis 13	0.67	0.15	Cell membrane 13	0.79	0.66
Microcystis 14	0.76	0.61	Cell membrane 14	0.88	0.83
Mean	0.66	0.36	Mean	0.76	0.65

objects separable in truth. Under this scheme, the lower Measure II, the better model in predictive power. Figure 3 illustrates the examples of Measure II.

3.2.2. Experiment Data. We compare the Mask R-CNN, weighted Mask R-CNN, and Mask Encoding for Single Shot Instance Segmentation (MEInst; [33]) models via Measure I and present the mean and standard errors given the prespecified number of circles and ellipses (i.e., 4 and 6) in Tables 1 and 2. The results indicate that the predicted mask of the weighted Mask R-CNN model is superior across simulation scenarios when we estimate the ground truth mask compared to the Mask R-CNN and MEInst. We train on the ResNet-50-FPN model as the backbone implemented in the PyTorch package for both the MEInst and weighted Mask R-CNN. The Mask RCNN runs at 67.47 ms per image with almost the same as the weighted Mask R-CNN records, and MEInst runs at 77.69 ms per image using our workstation (Intel i7-7800X, RAM 128GB, Geforce GTX 1080 Ti GPUs).

3.2.3. Real Data. In Table 3, we compare the performance of the Mask R-CNN and weighted Mask R-CNN models in real data. In algae data, mAP50 and Recall50 are 0.862 and 0.945 in the Mask R-CNN and 0.925 and 0.964 in the weighted Mask R-CNN, where mAP50 and Recall50 refer to mean AP and recall under IoU threshold of 0.5. In the same manner, mIoU in the Mask R-CNN is 0.801, and in the weighted Mask R-CNN, it is 0.845. In cell membrane data, mAP50 and Recall50 are 0.899 and 0.970 in the Mask R-CNN and 0.945 and 0.986 in the weighted Mask R-CNN. As a whole, it is evident that the weighted Mask R-CNN performs better than the Mask R-CNN in both microalgae and cell membrane data.

Furthermore, in Table 4, the comparisons in detecting borders between two models are given. We choose 14 algae images and 14 cell membrane images each. Hence, we can evaluate the area of images under the following conditions. First, the objects in images are detected in both the Mask R-CNN and weighted Mask R-CNN models. Second, the masks inferred from the Mask R-CNN are overlapped. This is reasonable in the sense that most of microalgae in an image are jumbled and many have put efforts to separating individual algae in vision to facilitate counting. Third, the specified objects are taxonomized as different groups. In Table 3, we observe that the weighted Mask R-CNN (i.e., the mean of 0.36 and 0.65 for algae and cell membrane) consistently outperforms the Mask R-CNN (i.e., the mean of 0.66 and 0.76 for algae and cell membrane) in separating boundaries of adjacent objects with respect to all target images. See the supplementary material (available here) for additional results.

4. Discussion

In this paper, we introduce the weighted Mask R-CNN specially designed to accurately segment instances. Simply put, this method accommodates in theory *a priori* known knowledge of boundary information in the midst of multiple objects. In numerical experiments, it is shown that the weighted Mask R-CNN model performs better than the Mask R-CNN and MEInst models in the boundary detection as stated in Tables 1 and 2. However, it is shown in the experiment that δ is required to be tuned properly to improve performance. In particular, we hardly perform the clear-cut for algae (e.g., *Microcystis*) and cell membrane images, in the sense that they are commonly mingled in an image and are formed with heterogeneous figures. To overcome this, the weighted Mask R-CNN is worth to implement the precise segmentation tasks. On top of that, it is also noteworthy that the proposed method can advance in microalgae research domain in keeping with improving instance segmentation. Surprisingly, this technique obviously contributes to quantify each single cell in vision sensing approaches. In reality, there are urgent needs in freshwater analysis to quantify the number of algae cells and the concentration of algae. This utility enables to monitor water quality in seas or rivers [34]. When

it comes to the model configuration, the weight in the model only builds on distance basis between objects, but yet this weight can be extended to other known knowledge in spirit of data integration. It is also interesting to exploit cutting-edge network architectures and modules in improving accuracy and accelerating the computational speed. We leave this subject for future study.

Data Availability

All data sets are available at the author's website (http://www.hifiai.pe.kr).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by Konkuk University in 2019. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning and Konkuk University Researcher Fund in 2020 and the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2019R1C1C1011366 and 2020R1C1C1A01005229).

Supplementary Materials

Figure S1: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S2: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S3: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S4: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S5: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S6: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S7: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S8: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S9: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S10: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S11: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). Figure S12: the instance segmentation examples of Mask R-CNN (left) and weighted Mask R-CNN (right). *(Supplementary Materials)*

References

[1] J. Ruiz-Santaquiteria, G. Bueno, O. Deniz, N. Vallez, and G. Cristobal, "Semantic versus instance segmentation in microscopic algae detection," *Engineering Applications of Artificial Intelligence*, vol. 87, article 103271, 2020.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[3] K. Schulze, U. M. Tillich, T. Dandekar, and M. Frohme, "PlanktoVision-an automated analysis system for the identification of phytoplankton," *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–10, 2013.

[4] G. Bueno, O. Deniz, A. Pedraza et al., "Automated diatom classification (part A): handcrafted feature approaches," *Applied Sciences*, vol. 7, no. 8, p. 753, 2017.

[5] J. Park, H. Lee, C. Y. Park, S. Hasan, T. Y. Heo, and W. H. Lee, "Algal morphological identification in watersheds for drinking water supply using neural architecture search for convolutional neural network," *Water*, vol. 11, no. 7, p. 1338, 2019.

[6] N. J. Poulton, "FlowCam: quantification and classification of phytoplankton by imaging flow cytometry," in *Imaging Flow Cytometry. Methods in Molecular Biology, vol 1389*, N. Barteneva and I. Vorobjev, Eds., Humana Press, New York, NY, USA, 2016.

[7] H. Zheng, H. Zhao, X. Sun, H. Gao, and G. Ji, "Automatic setae segmentation from *Chaetoceros* microscopic images," *Microscopy Research and Technique*, vol. 77, no. 9, pp. 684–690, 2014.

[8] N. Tang, F. Zhou, Z. Gu, H. Zheng, Z. Yu, and B. Zheng, "Unsupervised pixel-wise classification for *Chaetoceros* image segmentation," *Neurocomputing*, vol. 318, pp. 261–270, 2018.

[9] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 679–698, 1986.

[10] A. Victoria Anand Mary and G. Prabakaran, "An Efficient Automated Deep Learning Model For Diatom Image Segmentation And Classification," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, no. 11, pp. 446–454, 2019.

[11] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016, http://arxiv.org/abs/1602.07261.

[12] A. C. Jalba, M. H. Wilkinson, and J. B. Roerdink, "Automatic segmentation of diatom images for classification," *Microscopy Research and Technique*, vol. 65, no. 1-2, pp. 72–85, 2004.

[13] B. Rasti, D. Hong, R. Hang et al., "Feature extraction for hyperspectral imagery: the evolution from shallow to deep," 2020, http://arxiv.org/abs/2003.02822.

[14] D. Hong, L. Gao, N. Yokoya et al., "More diverse means better: multimodal deep learning meets remote-sensing imagery classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2020.

[15] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8246–8257, 2020.

[16] Z. Xie, U. K. Haritashya, V. K. Asari, B. W. Young, M. P. Bishop, and J. S. Kargel, "GlacierNet: a deep-learning approach for debris-covered glacier mapping," *IEEE Access*, vol. 8, pp. 83495–83510, 2020.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[18] X. Xia, C. Persello, and M. Koeva, "Deep fully convolutional networks for cadastral boundary detection from UAV images," *Remote Sensing*, vol. 11, no. 14, article 1725, 2019.

[19] L. Drăguț, O. Csillik, C. Eisank, and D. Tiede, "Automated parameterisation for multi-scale image segmentation on multiple layers," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 88, no. 100, pp. 119–127, 2014.

[20] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351*, N. Navab, J. Hornegger, W. Wells, and A. Frangi, Eds., pp. 234–241, Springer, Cham, 2015.

[22] E. K. Wang, C. M. Chen, M. M. Hassan, and A. Almogren, "A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain," *Future Generation Computer Systems*, vol. 108, pp. 135–144, 2020.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Caesar's Palace in Las Vegas, Nevada, USA, 2016.

[25] T. Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., pp. 740–755, Springer, Cham, 2014.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Hynes Convention Center in Boston, Massachusetts, USA, 2015.

[27] L. Weng, *Object Detection for Dummies Part 3: R-CNN Family*, Lilianweng. Gihub. Io/lil-log, 2017.

[28] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Las Condes Araucano Park, Región Metropolitana, Chile, 2015.

[29] F. Massa and R. Girshick, "maskrcnn-benchmark: Fast, Modular Reference Implementation of Instance Segmentation and Object Detection Algorithms in PyTorch," Mask R-CNN in PyTorch, 2018, Accessed: Apr, 29, 2019 https://github.com/facebookresearch/maskrcnn-benchmark.

[30] R. G. Wetzel and G. E. Likens, *Limnological Analyses*, Springer-Verlag, New York, NY, USA, 2 edition, 1991.

[31] J. K. Shin and K. J. Cho, "Distribution and population dynamics of *Microcystis* (Cyanophyta) in the Naktong River," *Algae*, vol. 12, no. 4, pp. 283–2903, 1997.

[32] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," *Advances in Neural Information Processing Systems*, vol. 25, pp. 2843–2851, 2012.

[33] R. Zhang, Z. Tian, C. Shen, M. You, and Y. Yan, "Mask encoding for single shot instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10226–10235, Washington State Convention Center in Seattle, Washington, USA, 2020.

[34] J. Park, D. Wang, and W. H. Lee, "Evaluation of weir construction on water quality related to algal blooms in the Nakdong River," *Environmental Earth Sciences*, vol. 77, no. 11, p. 408, 2018.

[35] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *ICCV*, vol. 1, no. 6, 2017.

[36] Q. Zhou, W. Chen, H. Zhang et al., "A flow cytometer based protocol for quantitative analysis of bloom-forming cyanobacteria (*Microcystis*) in lake sediments," *Journal of Environmental Sciences*, vol. 24, no. 9, pp. 1709–1716, 2012.

[37] A. Pedraza, G. Bueno, O. Deniz, G. Cristóbal, S. Blanco, and M. Borrego-Ramos, "Automated diatom classification (part B): a deep learning approach," *Applied Sciences*, vol. 7, no. 5, p. 460, 2017.

[38] A. Pedraza, G. Bueno, O. Deniz et al., "Lights and pitfalls of convolutional neural networks for diatom identification," in *Proceedings Volume 10679, Optics, Photonics, and Digital Technologies for Imaging Applications V*, Strasbourg, France, May 2018.

[39] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, http://arxiv.org/abs/1409.1556.