Biological Instance Segmentation with a Superpixel-Guided Graph

Xiaoyu Liu[1] , **Wei Huang**[1] , **Yueyi Zhang**[1,2] , **Zhiwei Xiong**[1,2,*]
[1]University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{liuxyu, weih527}@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

Abstract

Recent advanced proposal-free instance segmentation methods have made significant progress in biological images. However, existing methods are vulnerable to local imaging artifacts and similar object appearances, resulting in over-merge and over-segmentation. To reduce these two kinds of errors, we propose a new biological instance segmentation framework based on a superpixel-guided graph, which consists of two stages, *i.e.*, superpixel-guided graph construction and superpixel agglomeration. Specifically, the first stage generates enough superpixels as graph nodes to avoid over-merge, and extracts node and edge features to construct an initialized graph. The second stage agglomerates superpixels into instances based on the relationship of graph nodes predicted by a graph neural network (GNN). To solve over-segmentation and prevent introducing additional over-merge, we specially design two loss functions to supervise the GNN, *i.e.*, a repulsion-attraction (RA) loss to better distinguish the relationship of nodes in the feature space, and a maximin agglomeration score (MAS) loss to pay more attention to crucial edge classification. Extensive experiments on three representative biological datasets demonstrate the superiority of our method over existing state-of-the-art methods. Code is available at https://github.com/liuxy1103/BISSG.

1 Introduction

Instance segmentation aims at assigning a unique ID to each object in an image, which plays an important role in biological image analysis, such as morphology, distribution, and phenotyping [Minervini *et al.*, 2016]. Compared to natural images, instance segmentation for biological images is a more challenging task. On the one hand, the imaging quality of biological images is easily affected by sample collection and preparation, which leads to local imaging artifacts. On the other hand, the size, shape, and morphology of instances may vary greatly within an image.

Current learning-based methods can be mainly divided into proposal-based and proposal-free categories. The latter has attracted increasing research attention in biological images. Proposal-free methods predict well-designed instance-aware features and morphology properties, followed by a post-processing algorithm to yield final results. For example, [De Brabandere *et al.*, 2017] predicts instance-aware embeddings based on metric learning. [Liu *et al.*, 2018] analyzes the instance morphology characteristics. However, existing methods only focus on local features to separate adjacent and overlapping objects but ignore the object-level global context, which results in over-merge due to suffering local imaging artifacts. In addition, they rely heavily on the post-processing algorithm to cluster pixels into instances, which makes the clustered results prone to over-segmentation.

In this paper, we propose a new biological instance segmentation framework to address the above-mentioned problems. Distinct from existing proposal-free methods, our proposed framework is based on a superpixel-guided graph, which consists of two stages to reduce over-merge and over-segmentation, respectively. In the first stage, we utilize a U-Net to predict instance-aware embeddings and an affinity map to generate superpixels by the seeded watershed transformation. To avoid over-merge and make sure the superpixels are well-adhered to the instance boundary, the number of initial superpixels is controlled by distance transformation. We then convert these superpixels into a region adjacent graph (RAG) with the initial node and edge features extracted by the U-Net. In the second stage, a GNN consisting of multiple edge label graph neural network (EGNN) layers [Kim *et al.*, 2019] is used for edge classification to agglomerate superpixels into instances and solve over-segmentation. Specifically, the GNN aggregates graph features to capture global structure information of the graph and predicts the relationship of graph nodes.

Based on the proposed framework, we further design two loss functions to supervise the GNN. The repulsion-attraction (RA) loss is proposed to better distinguish the relationship of nodes representing superpixels in the feature space, which is different from the discriminative loss [De Brabandere *et al.*, 2017] used for pixel-level objects. As discussed in [Turaga *et al.*, 2009], a maximin edge in an undirected edge-weighted RAG is the highest minimal edge over all paths connecting a pair of nodes and is crucial to determine the relationship of nodes. Inspired by that, we design a maximin agglomeration

*Contact Author

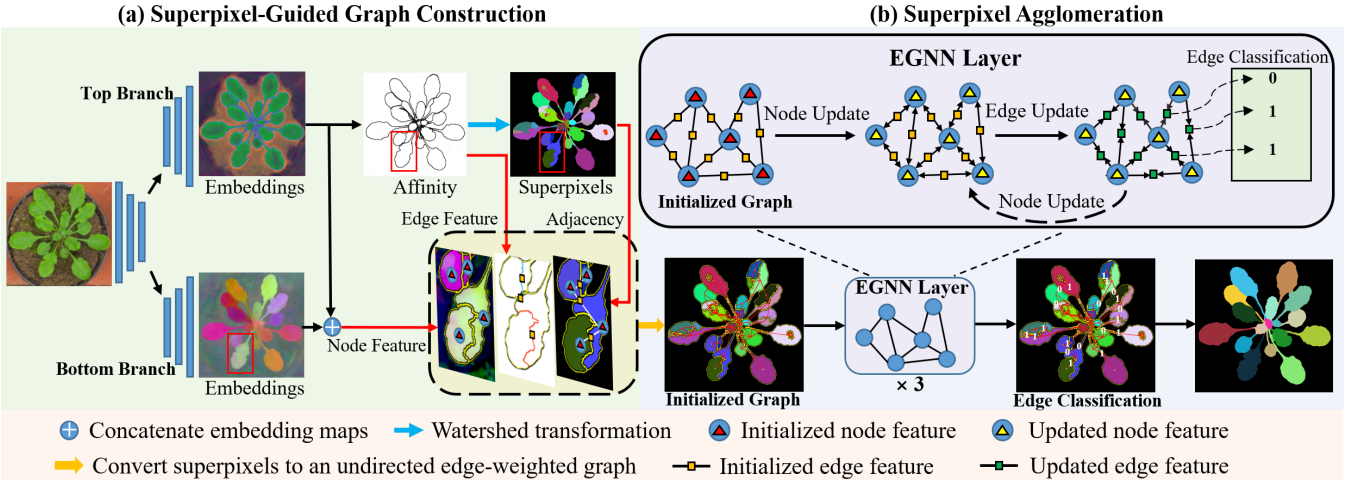

Figure 1: The framework of our proposed instance segmentation method for biological images. It consists of two stages: (a) Superpixel-Guided Graph Construction. It first predicts an instance-aware embedding map and an affinity map by a residual U-Net with dual decoders, from which node and edge features are extracted to construct an initialized graph. (b) Superpixel Agglomeration. It updates node and edge features to yield final segmentation by a GNN consisting of three cascaded EGNN layers.

score (MAS) loss to impose edge classifiers on paying more attention to the classification accuracy of the maximin edges. As demonstrated, these two loss functions significantly improve the segmentation performance.

2 Related Work

2.1 Proposal-Based Instance Segmentation

Proposal-based methods [He *et al.*, 2017; Liu *et al.*, 2019a; Zhang *et al.*, 2018] combine an object detection head to localize each object with a bounding box and an object segmentation head to predict foreground masks within each bounding box. Their performances are heavily reliant upon the prediction accuracy of bounding boxes, especially for biological images. On the one hand, it is difficult to distinguish adjacent instances by bounding boxes due to severe overlap among them. On the other hand, the size of instances, like neurons in Electron Microscopy (EM) images, is often larger than the receptive field of the model, which makes it impossible to use bounding boxes to locate a complete instance.

2.2 Proposal-Free Instance Segmentation

Proposal-free methods contain two main subcategories, *i.e.*, embeddings-based and affinity-based. Embedding-based methods [De Brabandere *et al.*, 2017; Kulikov and Lempitsky, 2020] predict high dimensional embeddings by a U-Net with a discriminative loss and then cluster the predicted embeddings into instances by a Mean-Shift algorithm [Fukunaga and Hostetler, 1975]. However, it is difficult to distinguish adjacent objects with similar morphology and appearances in the feature space. In addition, pixel-level features are sensitive to local imaging artifacts. Affinity-based methods [Beier *et al.*, 2017; Funke *et al.*, 2018; Liu *et al.*, 2018; Huang *et al.*, 2022] predict an affinity map that is similar to a pixel-wise boundary map, followed by a watershed or graph cut [Beier *et al.*, 2017] algorithm to cluster pixels into instances. However, affinity-based methods are susceptible to ambiguous boundaries, which results in segmentation errors. Moreover, the post-processing algorithms used in the above two kinds of methods are prone to over-segmentation. Especially for the neurons in EM images, the clustered results generated by the watershed transformation need to be refined by additional agglomeration algorithms to reduce over-segmentation. In contrast, we fuse the instance information from both embeddings and affinities, where embeddings and affinities are used to extract node and edge features, respectively. Therefore, our method is robust to local imaging artifacts and improves the distinguishability of adjacent objects in the feature space.

3 Proposed Method

3.1 Superpixel-Guided Graph Construction

As shown in Figure 1(a), a residual U-Net with an encoder and a dual-branch decoder (*i.e.* top and bottom branches with the same architecture) is adopted to extract the initial node and edge features. The top branch supervised by a binary cross-entropy (BCE) loss predicts an affinity map, which is converted to superpixels by the seeded watershed transformation. We control the number of superpixels based on enough seeds generated from the distance transformation. This is to avoid over-merge and make sure superpixels are well-adhered to the instance boundary. Following [De Brabandere *et al.*, 2017], the bottom branch predicts instance-aware embeddings supervised by a discriminative loss. This loss imposes pixels belonging to different instances to be discriminative in the feature space.

The above generated superpixels are then converted into a RAG by viewing each superpixel as a node and adding edges among all adjacent superpixels. Since affinities indicate instance boundaries, the embeddings generated from the penultimate layer in the top branch provide additional instance information to compensate for the embeddings generated in the bottom branch. Therefore, the feature of each node is built by

calculating the mean embeddings of the corresponding superpixel on the concatenated embedding maps of two branches. The edge feature is built by calculating the mean affinity intensities over the contour separating two adjacent superpixels. The influence of the number of superpixels on performance is discussed in Sec. 4.3.

3.2 Superpixel Agglomeration

As shown in Figure 1(b), we construct a GNN to agglomerate superpixels into segments by predicting the relationships of nodes. The GNN consists of three EGNN layers to avoid the over-smoothing issue. Each EGNN layer contains a node feature update block and an edge feature update block.

Let $G = (V, E)$ be an undirected weighted graph with N nodes. $V = \{v_i\}$ is the set of nodes and $E = \{e_{ij}|v_i : v_j\}$ is the set of edges, where $1 \leq i, j \leq N$ and v_i is one of the first-order or higher-order neighboring nodes of v_j. $X \in \mathbb{R}^{N\times F}$ represents the node feature matrix of the whole graph, where $x_i = X[i,:]^T$ is a F-dimensional feature vector of the node $v_i \in V$. The agglomeration procedure can be regarded as a binary edge classification problem by predicting the relationship of adjacent superpixels. Therefore, we define an edge feature tensor $A \in \mathbb{R}^{N\times N\times 2}$, which is a variant of the adjacent matrix. We initialize the input edge feature $A^0_{ij} = \left[A^0_{ij0} \,||\, A^0_{ij1}\right]$ of the first EGNN layer as

$$A^0_{ij} = \begin{cases} [1 - \hat{a_{ij}} \,||\, \hat{a_{ij}}] & \text{if } e_{ij} \in E \\ [0||0] & \text{if } e_{ij} \notin E \end{cases}, \tag{1}$$

where $||$ is the concatenation operation, $\hat{a_{ij}}$ represents the mean affinity intensity over the contour of two adjacent superpixels v_i and v_j. A_{ij1} is the agglomeration score which represents the probability of merging two adjacent nodes, and A_{ij0} is the probability of non-merging.

The update of node features is utilized to aggregate features of neighboring nodes, which captures the structural information of graph. An update of a node feature is formulated as

$$x^l_i = f_n\Big(\sum_{e_{ij}\in E} A^{l-1}_{ij0} x^{l-1}_j \,||\, \sum_{e_{ij}\in E} A^{l-1}_{ij1} x^{l-1}_j \,||\, x^{l-1}_i\Big), \tag{2}$$

where x^l_i represents a node feature on the l^{th} EGNN layer, and f_n is a multi-layer perceptron.

The newly updated node features are used for updating edge features by estimating the feature similarity s_{ij} between two adjacent nodes. The edge feature A_{ij} is updated as

$$\begin{aligned} A^l_{ij} &= \begin{cases} [(1 - s_{ij})A^{l-1}_{ij0} \,||\, s_{ij}A^{l-1}_{ij1}] & \text{if } e_{ij} \in E \\ [0||0] & \text{if } e_{ij} \notin E \end{cases}, \\ s_{ij} &= f_e(x^{l-1}_i - x^{l-1}_j \,||\, \hat{a_{ij}}), \end{aligned} \tag{3}$$

where f_e is a multi-layer perceptron activated by a sigmoid function.

Following [Kim *et al.*, 2019], the updated edge features are normalized to ensure that each element of A^l is in the interval [0, 1]. Each EGNN layer outputs an updated edge feature tensor A^l. We calculate the mean scores predicted by all EGNN layers to agglomerate the superpixels.

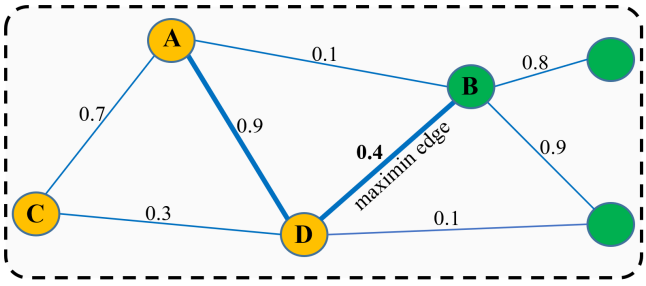

Figure 2: A simple schematic diagram to show the maximin edge between node A and node B in the whole graph. We search for minimal edges in each path connecting the pair of nodes and then select the maximal one among these minimal edges as the maximin edge. The path (A-D-B) containing the maximin edge is highlighted by the bold line. Different node colors indicate different instances. Values on edges refer to the predicted agglomeration scores.

3.3 RA Loss and MAS Loss

RA loss and MAS loss are proposed to supervise the learning of node and edge features in the stage of superpixel agglomeration, respectively. To better describe these two loss functions, we first define the edge classification label: $\delta(i, j) = 1$, if the adjacent nodes v_i and v_j belong to the same instance. $\delta(i, j) = 0$, if the nodes v_i and v_j belong to different instances or are not adjacent.

If superpixels belong to different instances, we expect their node features to be far away in the feature space. On the contrary, we expect that node features from the same instance are close. Different from the discriminative loss [De Brabandere *et al.*, 2017] as a pixel-level constraint, the RA loss is designed to discriminate node features. As shown in Figure 2, the RA loss makes the node features of different instances repel each other (nodes A and B), while the node features of the same instances attract each other (nodes A and C). We formulate the RA loss as

$$\begin{aligned} L_{RA} &= \sum_{i=1}^{N}\sum_{j=1}^{N} l_{RA}(i, j), \\ l_{RA}(i, j) &= \begin{cases} [2d - \|x_i - x_j\|]^2_+ & \text{if } \delta(i, j) = 0 \\ [\|x_i - x_j\| - 2d]^2_+ & \text{if } \delta(i, j) = 1 \end{cases}, \end{aligned} \tag{4}$$

where $\|\cdot\|$ represents L_2 distance, $[x]_+ = max(0, x)$ denotes the hinge, and d is the margin distance of two node features from different instances. We use $d = 1.5$ here, which is the same as the margin in [De Brabandere *et al.*, 2017].

As demonstrated in [Turaga *et al.*, 2009], edge classification results with low error rates still lead to poor agglomeration results when supervised by a simple BCE loss. Whether a pair of adjacent superpixels are merged depends on the maximin edge of the two corresponding nodes in the whole graph. A maximin edge is the highest minimal edge over all paths connecting the pair of nodes. To find it, we adopt a maximum spanning tree to search for minimal edges in each path connecting the pair of nodes and then select the maximal one among these minimal edges as the maximin edge. Therefore, the MAS loss is designed to supervise the learning of edge features to optimize global results. As shown in Figure 2,

Method	**SBD** ↑	|DiC| ↓
Nottingham[Scharr *et al.*, 2016]	68.3	3.8
IPK[Pape and Klukas, 2014]	74.4	2.6
AC[Araslanov *et al.*, 2019]	79.1	1.1
Discriminative[De Brabandere *et al.*, 2017]	84.2	1.0
Recurrent attention[Ren and Zemel, 2017]	84.9	**0.8**
Data augmentation[Kuznichov *et al.*, 2019]	88.7	5.4
Harmonic[Kulikov and Lempitsky, 2020]	89.9	3.0
Synthesis data[Ward *et al.*, 2018]	90.0	-
PFFNet[Liu *et al.*, 2021]	91.1	-
Discri. w/ our emb.	88.2	1.6
Ours	**91.7**	1.4

Table 1: Quantitative comparison with state-of-the-art methods on the test set of CVPPP A1.

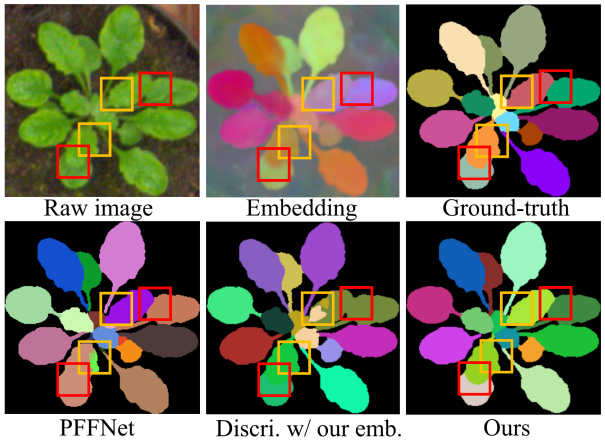

Figure 3: Visual results on the CVPPP dataset. Over-merge and over-segmentation in 'Discriminative' and 'PFFNet' are highlighted by red and gold boxes.

Dataset	Method	VOI-s ↓	Arand-s ↓	**VOI** ↓	ARAND ↓
AC3/AC4	MALA	1.4759	0.4392	0.8730	0.2222
	LMC	1.5398	0.4749	0.8969	0.2397
	Ours	1.5509	0.4790	**0.8181**	**0.2059**
FIB-25	MALA	1.7846	0.3811	1.4027	0.2485
	LMC	1.8284	0.3916	1.4842	0.2524
	Ours	1.8566	0.3981	**1.3450**	**0.2467**
CREMI	MALA	1.6262	0.4471	0.9368	0.1922
	LMC	1.5697	0.4211	0.9602	0.1940
	Ours	1.6381	0.4549	**0.9141**	**0.1867**

Table 2: Quantitative comparison on three representative EM datasets. 'VOI-s' and 'Anrand-s' represent the intermediate superpixels results in the first stage.

we consider all paths connecting node A and node B to find a maximin edge, based on these agglomeration scores predicted by the GNN. The MAS loss is formulated as

$$
\begin{aligned}
L_{MAS} &= \sum_{e_{ij} \in E} (mm(i,j) - \delta(i,j))^2, \\
p^*_{i,j} &= \arg\max_{p \in P_{i,j}} \min_{e_{mn} \in p} A_{mn1}, \\
mm(i,j) &= \min_{e_{mn} \in p^*_{i,j}} A_{mn1},
\end{aligned}
\tag{5}
$$

where $p^*_{i,j}$ is the path including a maximin edge among all paths connecting nodes v_i and v_j, and $mm(i,j)$ is the predicted agglomeration score of the maximin edge.

3.4 Training Strategy

To improve the training efficiency of our framework, the training procedure is divided into two phases. We pre-train the U-Net for 200 epochs during the first phase. Then we train the GNN for 200 epochs during the second phase. We update the parameters of the GNN and the bottom decoder branch of the U-Net synchronously, but the parameters of the encoder and the top decoder branch of the U-Net are frozen in the second phase. Each EGNN layer is supervised by the RA loss and the MAS loss to update node and edge features.

The initial learning rates are set as 10^{-4} and 10^{-3} for U-Net and GNN respectively, decayed by a half when the loss stops improving in 30 epochs. The batch size is set as 4 and 1 for the first and second phase. The Adam optimizer is adopted during the training with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. Experiments are implemented on one NVIDIA TitanXP GPU. The way of back-propagation is further discussed in Sec. 4.3.

4 Experiments

4.1 Datasets and Metrics

Plant Phenotype Images

The CVPPP A1 dataset [Scharr *et al.*, 2014] is one of the most popular instance segmentation benchmarks, which consists of 128 training images and 33 testing images with a resolution of 530×500. We randomly select 20 images from the training set as the validation set. This dataset is challenging due to the wide variety of leaf shapes and severe occlusion among leaves. The quality of the segmentation result is measured by Symmetric Best Dice (SBD) and absolute Difference in Counting (|DiC|) metrics.

Electron Microscopy Images

AC3/AC4 [Kasthuri *et al.*, 2015], CREMI, and FIB-25 [Takemura *et al.*, 2015] are three popular Electron Microscopy (EM) datasets. The AC3/AC4 dataset consists of two human-labeled sub-volumes imaged from the mouse somatosensory cortex. These two sub-volumes contain 256 and 100 images (1024×1024), respectively. We use the top 226 slices of AC3 for training, the rest 30 slices for validation, and AC4 for testing. The CREMI dataset is composed of three sub-volumes imaged from the adult drosophila brain, and each contains 125 images (1250×1250). We use the top 50 slices for testing, the middle 60 slices for training, and the bottom 15 slices for validation. The FIB-25 dataset is also imaged from a drosophila brain, which is composed of two sub-volumes with a resolution of $520 \times 520 \times 520$. We use one sub-volume for training, in which the bottom 50 slices are used for validation, and use the other sub-volume for testing. We adopt two widely used metrics in the field of EM image segmenta-

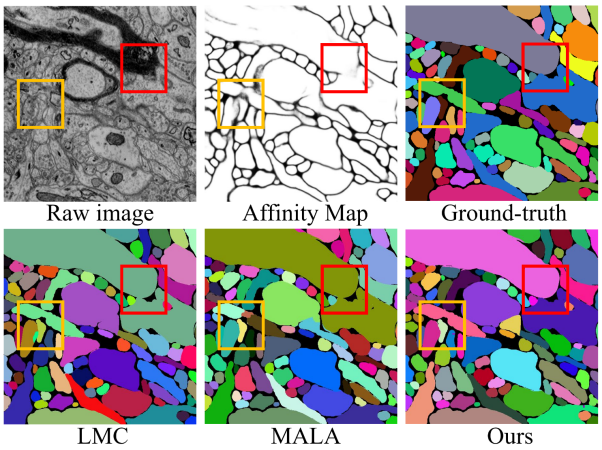

Figure 4: Visual results on the AC3/AC4 datasets. Over-merge and over-segmentation in 'LMC' and 'MALA' are highlighted by red and gold boxes, respectively.

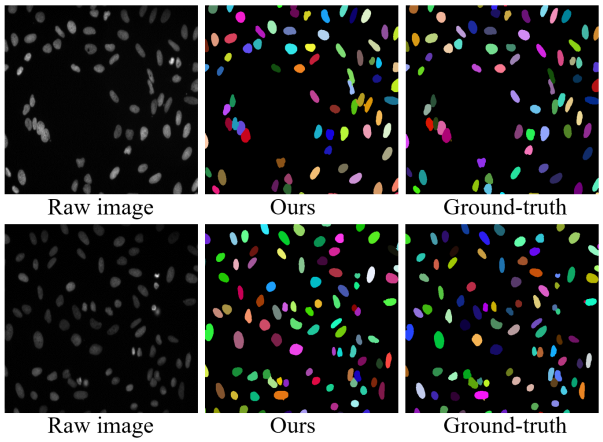

Figure 5: Visual results on the BBBC039 dataset.

Method	AJI ↑	Dice ↑	F1 ↑	**PQ** ↑
Mask RCNN[He *et al.*, 2017]	0.7983	0.9277	0.9180	0.7773
Cell RCNN[Zhang *et al.*, 2018]	0.8070	0.9290	0.9276	0.7959
UPSNetN[Xiong *et al.*, 2019]	0.8128	0.9274	0.9191	0.7857
JSISNet[De Geus *et al.*, 2018]	0.8134	0.9316	0.9282	0.7913
PanFPN[Kirillov *et al.*, 2019]	0.8193	0.9320	0.9275	0.7960
OANet[Liu *et al.*, 2019b]	0.8198	0.9372	0.9330	0.8085
AUNet[Li *et al.*, 2019]	0.8252	0.9377	0.9315	0.8090
Cell RCNNv2[Liu *et al.*, 2019a]	0.8260	0.9336	0.9328	0.8010
PFFNet[Liu *et al.*, 2021]	0.8477	0.9478	0.9451	0.8331
Ours	**0.8680**	**0.9482**	**0.9670**	**0.8629**

Table 3: Quantitative comparison with state-of-the-art methods on the test set of BBBC039V1.

tion for quantitative evaluation, *i.e.,* Variation of Information (VOI) and Adapted Rand Error (ARAND).

Fluorescence Microscopy Images

The BBBC039V1 dataset [Ljosa *et al.*, 2012] contains 200 images (520×696) from Fluorescence Microscopy (FM). The objects in each image are cells with various shapes and densities. Following the official data split, we use 100 images for training, 50 images for validation, and the rest 50 images for testing. For evaluation, we adopt four common metrics for cell segmentation in FM images, *i.e.,* Aggregated Jaccard Index (AJI), object-level F1 score (F1), Panoptic Quality (PQ), and pixel-level Dice score (Dice).

4.2 Experimental Results

Plant Phenotype Images

We compare our method with existing methods on the test set of CVPPP A1. As shown in Table 1, our method achieves the best results on the SBD metric, which is the key performance indicator of this dataset. Meanwhile, compared with the most recent methods, we achieve competitive results on the |DiC| metric (early methods with the best |DiC| results are far behind on SBD). It is worth mentioning that, if we directly adopt the Mean-Shift algorithm to cluster the predicted embeddings in the first stage of our framework, the result can be regarded as an enhanced baseline for 'Discriminative' [De Brabandere *et al.*, 2017] called 'Discri. w/ our emb.' We further visualize the intermediate embeddings and the affinity map in our first stage, and qualitatively compare them with baseline methods, as shown in Figure 3. We can observe that: (1) 'PFFNet' as a proposal-based method is limited by the prediction of bounding boxes. Leaves with challenging occlusions are difficult to be approximated by bounding boxes. (2) 'Discri. w/ our emb.' directly clusters embeddings into instances, which is prone to over-merge due to locally similar features, especially for adjacent objects. (3) Our method makes up for these weaknesses by combining instance information from both embeddings and affinities, and capturing global structure information with a GNN, which significantly reduces over-merge and over-segmentation.

Electron Microscopy Images

We compare our method with two advanced methods, *i.e.*, MALA [Funke *et al.*, 2018] and LMC [Beier *et al.*, 2017] on three EM datasets. Both MALA and LMC are two-stage segmentation methods and adopt different post-processing algorithms to agglomerate superpixels from the first stage. For a fair comparison, we adopt the same backbone of U-Net for the first stage as used in our method. As shown in Table 2, our method achieves the best quantitative results in terms of both VOI and ARAND on all datasets. Meanwhile, it reduces over-merge and over-segmentation compared with the baseline methods, as shown in Figure 4. An additional observation is that 'Affinity Map' provides useful boundary information to separate adjacent objects but suffers from certain boundary errors, which may lead to over-segmentation and over-merge. In the proposed framework, we generate enough superpixels to avoid over-merge in the first stage and then solve over-segmentation in the second stage.

Fluorescence Microscopy Images

We demonstrate the effectiveness of our method on the BBBC039V1 dataset. As listed in Table 3, our method outperforms existing methods on all metrics, significantly im-

Embedding1	Embedding2	Affiniy	**VOI** ↓	ARAND ↓
✓			1.1223	0.2802
	✓		0.9536	0.2405
✓	✓		0.9356	0.2366
✓		✓	0.8583	0.2305
	✓	✓	0.8478	0.2245
✓	✓	✓	**0.8181**	**0.2059**

Table 4: Ablation study on different parts of graph features. 'Embedding1' and 'Embedding2' represent the embeddings predicted by the top branch and the bottom branch of U-Net, respectively.

Decoder1	Decoder2	Encoder	**VOI** ↓	ARAND ↓
			0.8359	0.2263
✓			0.8619	0.2264
	✓		**0.8181**	**0.2059**
✓	✓		0.8499	0.2293
✓	✓	✓	0.8772	0.2348

Table 5: Ablation study on different implementations of back-propagation. 'Decoder1' and 'Decoder2' represent the top branch and the bottom branch of U-Net, respectively.

proving the key PQ metric by 3.6%. Visual results demonstrate the superior segmentation performance of our method in dealing with challenging cases, as shown in Figure 5.

4.3 Ablation Study

To investigate the effectiveness of different components in the proposed framework,we conduct a series of ablation studies on the AC3/AC4 dataset.

Graph Features

We conduct ablation experiments to validate the effectiveness of each part of the graph features extracted from the embeddings (from the top and bottom branches) and affinities predicted by the U-Net, as shown in Table 4. It suggests that the three parts of graph features are complementary to each other. The best result is achieved through their combination. The initial node features extracted from the bottom branch contain more effective information for decision-making than that from the top branch. The edge features from the mean affinities over the contour separating adjacent superpixels contain powerful priors of instances boundaries, which is beneficial to agglomerate superpixels.

Back-Propagation

In the second training phase, we explore different implementations of back-propagation. We choose different parts of U-Net to update with the GNN synchronously and freeze the other parameters. As shown in Table 5, the setting of back-propagation from the GNN to the bottom decoder branch achieves the best performance. The GNN loss and the discriminative loss on U-Net can promote each other to obtain node features. However, the GNN gradient update used for the top decoder branch and encoder of the U-Net degrades performance, mainly because the watershed transformation is not directly controlled by the network.

RA loss	MAS loss	BCE loss	**VOI** ↓	ARAND ↓
✓			0.9498	0.2450
	✓		0.8257	0.2153
		✓	0.8922	0.2228
✓		✓	0.8743	0.2510
✓	✓		**0.8181**	**0.2059**

Table 6: Ablation study on the proposed RA loss and MAS loss.

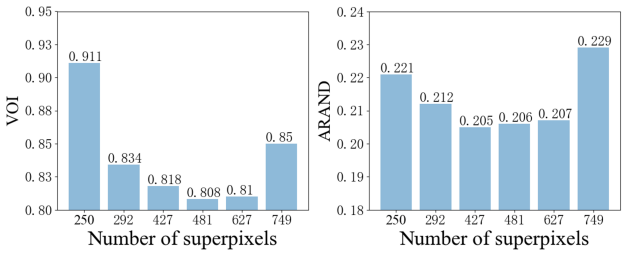

Figure 6: Ablation study on different numbers of superpixels generated by the watershed transformation.

RA Loss and MAS Loss

The comparison of the MAS loss and the simple BCE loss for binary edge classification is shown in Table 6. The model trained with the single MAS loss achieves significant improvement over the model trained with the single BCE loss. Moreover, the model with additional supervision by the RA loss for node features achieves better results. It validates that the RA loss plays an important role in distinguishing nodes in the feature space.

Number of Superpixels

The effect of the number of superpixels generated in the first stage is shown in Figure 6. Generating more superpixels can reduce over-merge in the first stage, which is beneficial for the final segmentation. However, when the number of superpixels gets too large and exceeds the receptive field of the GNN, the performance drops.

5 Conclusion

We propose a new biological instance segmentation framework based on a superpixel-guided graph, which significantly reduces over-merge and over-segmentation. The essential is introducing a GNN to model the relationship of superpixels, which is supervised by two specially designed loss functions. The RA loss better distinguishes the relationship of nodes in the feature space, and the MAS loss pays more attention to the crucial edge classification. Our proposed method achieves notably improved performance over existing state-of-the-art methods on three representative biological datasets.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2017YFA0700800, the National Natural Science Foundation of China under Grant 62021001, and the University Synergy Innovation Program of Anhui Province No. GXXT-2019-025.

References

[Araslanov *et al.*, 2019] Nikita Araslanov, Constantin A Rothkopf, and Stefan Roth. Actor-critic instance segmentation. In *CVPR*, 2019.

[Beier *et al.*, 2017] Thorsten Beier, Constantin Pape, Nasim Rahaman, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature Methods*, 14(2):101–102, 2017.

[De Brabandere *et al.*, 2017] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017.

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

[Fukunaga and Hostetler, 1975] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

[Funke *et al.*, 2018] Jan Funke, Fabian Tschopp, William Grisaitis, et al. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1669–1680, 2018.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[Huang *et al.*, 2022] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *AAAI*, 2022.

[Kasthuri *et al.*, 2015] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015.

[Kim *et al.*, 2019] Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. Edge-labeling graph neural network for few-shot learning. In *CVPR*, 2019.

[Kirillov *et al.*, 2019] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019.

[Kulikov and Lempitsky, 2020] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *CVPR*, 2020.

[Kuznichov *et al.*, 2019] Dmitry Kuznichov, Alon Zvirin, Yaron Honen, and Ron Kimmel. Data augmentation for leaf segmentation and counting tasks in rosette plants. In *CVPRW*, 2019.

[Li *et al.*, 2019] Yanwei Li, Xinze Chen, Zheng Zhu, et al. Attention-guided unified network for panoptic segmentation. In *CVPR*, 2019.

[Liu *et al.*, 2018] Yiding Liu, Siyu Yang, Bin Li, et al. Affinity derivation and graph merge for instance segmentation. In *ECCV*, 2018.

[Liu *et al.*, 2019a] Dongnan Liu, Donghao Zhang, Yang Song, Chaoyi Zhang, Fan Zhang, Lauren O'Donnell, and Weidong Cai. Nuclei segmentation via a deep panoptic model with semantic feature fusion. In *IJCAI*, 2019.

[Liu *et al.*, 2019b] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *CVPR*, 2019.

[Liu *et al.*, 2021] Dongnan Liu, Donghao Zhang, Yang Song, Heng Huang, and Weidong Cai. Panoptic feature fusion net: A novel instance segmentation paradigm for biomedical and biological images. *IEEE Transactions on Image Processing*, 30:2045–2059, 2021.

[Ljosa *et al.*, 2012] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature Methods*, 9(7):637–637, 2012.

[Minervini *et al.*, 2016] Massimo Minervini, Andreas Fischbach, Hanno Scharr, and Sotirios A Tsaftaris. Finely-grained annotated datasets for image-based plant phenotyping. *Pattern recognition letters*, 81:80–89, 2016.

[Pape and Klukas, 2014] Jean-Michel Pape and Christian Klukas. 3-d histogram-based segmentation and leaf detection for rosette plants. In *ECCV*, 2014.

[Ren and Zemel, 2017] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In *CVPR*, 2017.

[Scharr *et al.*, 2014] Hanno Scharr, Massimo Minervini, Andreas Fischbach, and Sotirios A Tsaftaris. Annotated image datasets of rosette plants. In *ECCV*, 2014.

[Scharr *et al.*, 2016] Hanno Scharr, Massimo Minervini, Andrew P French, et al. Leaf segmentation in plant phenotyping: a collation study. *Machine Vision and Applications*, 27(4):585–606, 2016.

[Takemura *et al.*, 2015] Shin-ya Takemura, C Shan Xu, Zhiyuan Lu, et al. Synaptic circuits and their variations within different columns in the visual system of drosophila. *Proceedings of the National Academy of Sciences*, 112(44):13711–13716, 2015.

[Turaga *et al.*, 2009] Srinivas C Turaga, Kevin L Briggman, Moritz Helmstaedter, Winfried Denk, and H Sebastian Seung. Maximin affinity learning of image segmentation. *arXiv preprint arXiv:0911.5372*, 2009.

[Ward *et al.*, 2018] Daniel Ward, Peyman Moghadam, and Nicolas Hudson. Deep leaf segmentation using synthetic data. *arXiv preprint arXiv:1807.10931*, 2018.

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

[Zhang *et al.*, 2018] Donghao Zhang, Yang Song, Dongnan Liu, Haozhe Jia, Siqi Liu, Yong Xia, Heng Huang, and Weidong Cai. Panoptic segmentation with an end-to-end cell r-cnn for pathology image analysis. In *MICCAI*, 2018.