Learning Cross-Representation Affinity Consistency for Sparsely Supervised Biomedical Instance Segmentation

Xiaoyu Liu[1,*] Wei Huang[1] Zhiwei Xiong[1,2,†]
Shenglong Zhou[1] Yueyi Zhang[1,2] Xuejin Chen[1,2] Zheng-Jun Zha[1] Feng Wu[1,2]

[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

Abstract

Sparse instance-level supervision has recently been explored to address insufficient annotation in biomedical instance segmentation, which is easier to annotate crowded instances and better preserves instance completeness for 3D volumetric datasets compared to common semi-supervision. In this paper, we propose a sparsely supervised biomedical instance segmentation framework via cross-representation affinity consistency regularization. Specifically, we adopt two individual networks to enforce the perturbation consistency between an explicit affinity map and an implicit affinity map to capture both feature-level instance discrimination and pixel-level instance boundary structure. We then select the highly confident region of each affinity map as the pseudo label to supervise the other one for affinity consistency learning. To obtain the highly confident region, we propose a pseudo-label noise filtering scheme by integrating two entropy-based decision strategies. Extensive experiments on four biomedical datasets with sparse instance annotations show the state-of-the-art performance of our proposed framework. For the first time, we demonstrate the superiority of sparse instance-level supervision on 3D volumetric datasets, compared to common semi-supervision under the same annotation cost. Code is available at https://github.com/liuxy1103/CRAC.

1. Introduction

Biomedical instance segmentation aims to assign each image pixel to an instance, which plays an essential role in biomedical instance morphology and distribution analysis [36, 32]. Recently, deep-learning-based methods [22, 26, 48, 20, 15, 30, 6] have achieved state-of-the-art performance on biomedical instance segmentation with densely annotated images for fully-supervised training. However, it is expensive and laborious to obtain accurate and dense annotations.

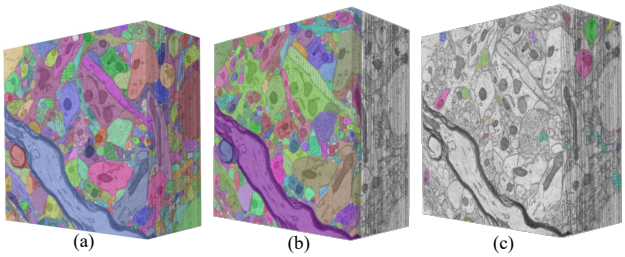

Figure 1. Comparison of instance annotation schemes in different supervision ways for the volumetric dataset. (a) Full supervision, where the whole volume is annotated. (b) Semi-supervision, where a subset of the whole volume is densely annotated. (c) sparse instance-level supervision, where a subset of instances of the whole volume is annotated.

To alleviate the huge cost of manual annotations, many works [42, 13, 49, 37, 7] adopt the semi-supervised learning strategy to train an effective segmentation model with a small set of densely annotated images and a large set of unannotated images. However, in the field of biomedical instance segmentation, dense annotation is difficult to apply due to the crowded instances. Especially, for the 3D volumetric datasets (*e.g.*, electron microscopy datasets), dense annotation within a small sub-volume usually damages the completeness of the instances along the axial direction. Meanwhile, in practice, domain experts tend to annotate instances one by one sparsely, rather than annotating all instances densely within a small sub-volume.

In this paper, we aim to deal with the task of biomedical instance segmentation with sparse instance annotations. This task is of utmost importance, as annotating a small subset of instances in a biomedical dataset not only saves the annotation cost but also conforms to annotation habits in the practical scenario. We illustrate different annotation

*This work was done during Xiaoyu Liu's internship at Institute of Artificial Intelligence, Hefei Comprehensive National Science Center.
†Corresponding author: zwxiong@ustc.edu.cn.

schemes in Fig. 1 and focus on the comparison between common semi-supervision and sparse instance-level supervision. For a volumetric dataset, a few consecutive images are densely annotated for the semi-supervision in Fig. 1 (b), which generates many incomplete instances along the axial direction. Instead, a few instances are randomly annotated for the sparse instance-level supervision in Fig. 1 (c), which preserves complete 3D structure information of each instance under the same number of annotated voxels as the semi-supervision.

Actually, sparse instance-level supervision can be considered a special kind of semi-supervision at the region level. Thus, existing semi-supervised methods [42, 13, 49, 37, 7] can be readily adapted to the sparse instance-level supervision. As a pioneer work, SPOCO [46] imposes a consistency regularization scheme [42] in the unannotated regions of two embedding maps predicted in different augmented views. However, SPOCO adopts a metric learning [9]-based segmentation method to distinguish different instances in the feature space, whose performance heavily drops as the number of annotated instances decreases. On the other hand, the perturbation consistency used in SPOCO is relatively simple and cannot effectively exploit instance structure information in unannotated regions. In other words, there remains a large room to release the potential of the sparse instance-level supervision.

Compared to metric learning, affinity-based instance segmentation methods [44, 43] are less sensitive to the number of annotated instances, since affinities encode instance boundary structure knowledge and are easier to be learned. Inspired by this kind of methods, we propose an effective biomedical instance segmentation framework with sparse instance annotations, which learns cross-representation affinity consistency. Specifically, our framework consists of two individual networks to predict two different representations, *i.e.*, an embedding map and an implicit affinity map (IAM), respectively. We then calculate the similarity between pixel embeddings and convert the embedding map into an explicit affinity map (EAM). The IAM exploits the spatial structure information of instances, while the EAM exploits the semantic information of instances in the feature space. By building the perturbation consistency between the IAM and the EAM, we combine the advantages of two kinds of affinity modeling to capture both feature-level instance discrimination and pixel-level instance boundary structure. This perturbation consistency is different from the existing perturbation consistency between the same kinds of representations widely used in semi-supervised learning.

In our framework, we use the groundtruth affinity map generated from sparse instance annotations to supervise the two affinity maps in the annotated regions. Inspired by the cross pseudo supervision [7], we further propose an affinity cross-supervision mechanism to facilitate affinity consistency learning on the unannotated regions. Specifically, we design a pseudo-label noise filtering scheme that integrates a heuristic decision strategy based on an adaptive threshold and a learning decision strategy based on a pre-trained confident pixel selection network (CPSN), to select highly confident regions as pseudo labels from one affinity map to supervise the other.

We evaluate our methods on four biomedical datasets with sparse instance annotations to demonstrate the state-of-the-art performance of our proposed framework, compared to three kinds of baselines: 1) the existing sparsely supervised method SPOCO, 2) advanced semi-supervised methods which are adapted to sparse instance-level supervision, and 3) original semi-supervised methods trained with dense annotation under the same annotation cost.

The contributions of this paper are as follows:

- We propose a sparsely supervised biomedical instance segmentation framework by enforcing the perturbation consistency between two kinds of affinity modeling.
- We propose an affinity cross-supervision mechanism with a pseudo-label noise filtering scheme integrating two decision strategies to select highly confident regions to facilitate affinity consistency learning.
- We conduct extensive experiments on four biomedical datasets with sparse instance annotations to demonstrate the state-of-the-art performance of our method.
- We demonstrate the general superiority of sparse instance-level supervision over semi-supervision on 3D volumetric datasets for the first time.

2. Related Work

2.1. Biomedical Instance Segmentation

The prevalent instance segmentation methods are mainly divided into two categories: proposal-based [12, 26, 3, 47] and proposal-free [5, 38, 27, 29, 28, 33]. The former is based on object detection to distinguish different instances, while the latter is based on instance-aware features and morphology properties, which is more suitable to complex and dense instances of biomedical datasets. Metric learning [9, 22] and affinity learning [11, 31] are two proposal-free representatives.

Metric learning [9, 22, 20] uses a discriminative loss [9] to impose pixels belonging to different instances to be discriminative from each other in the feature space. However, as the number of instances decreases in the sparse instance-level supervision setting, it is difficult for the network to distinguish different instances in the feature space.

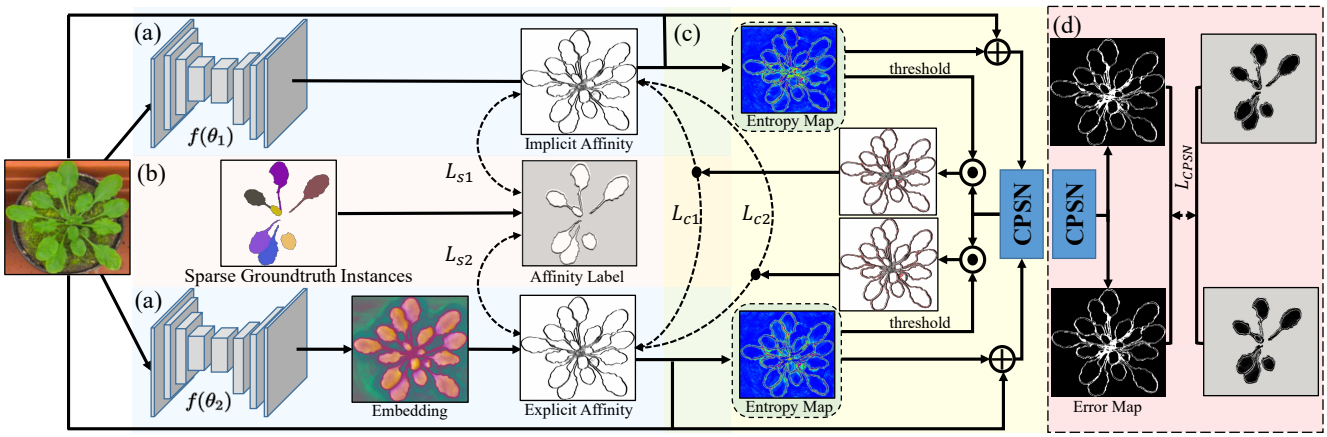

Figure 2. The workflow of our proposed sparsely supervised biomedical instance segmentation framework. (a) The input image is fed into two networks $f(\theta_1)$ and $f(\theta_2)$ to predict an affinity map and an embedding map, respectively. The embedding map is then explicitly converted into an affinity map. (b) These two affinity maps are supervised by the affinity label on the annotated region. The unannotated region of the affinity label is highlighted in gray color. The white and black colors indicate the affinity value as 1 and 0. (c) The highly confident region with low entropy of each affinity map is selected by a confident pixel selection network (CPSN) to supervise the other on the unannotated region. The red color indicates the region of the affinity map with high entropy, which is not used for cross-supervision. (d) This diagram illustrates the training stage of the CPSN, in which the model is trained on the annotated region to predict regions with segmentation errors. These regions are then excluded from being used as pseudo labels to supervise the other affinity map. $\odot$ and $\oplus$ represent dot product and concatenation operations, respectively.

Affinity learning [2, 11, 31, 14] predicts pixel affinities that only encode spatial structure information between adjacent instances without ensuring the uniqueness of each instance in the feature space. Thus, the affinities are easy to be learned by the network and less affected by the number of instances. Even if there is only one annotated instance, this method can learn affinities between the unannotated region and the instance.

2.2. Consistency Regularization

Consistency regularization [1] plays a vital role in semi-supervised learning. It enforces the consistency of predictions of unannotated data with various perturbations, by introducing a regularization loss function. The categories of perturbation are mainly divided into input perturbation [49, 25, 10], feature perturbation [37, 40], and network perturbation [7, 17]. Existing methods generally enforce the consistency between different perturbations of the same representation. Take an example of combining mean-teacher [42] and the affinity-based segmentation method: both the teacher and student networks directly predict affinity maps that belong to the same representation. Different from existing consistency regularization methods, our proposed framework enforces consistency between two different representations from two kinds of affinity modeling.

3. Problem Formulation

Given an input image I with a size of $H \times W$, containing K instances (including background), M of which are annotated ($M \ll K$). The image I is divided into two pixel sets, *i.e.*, an annotated pixel region R^L and an unannotated pixel region R^U. The unannotated region R^U contains unannotated instances and background. Different from semi-supervised learning, our goal is to train a model to predict the accurate affinity map by leveraging both a few annotated instances in the annotated region R^L and the unannotated region R^U.

Before introducing our method, we clarify the definition of affinity. The affinity map $A = [a_1, a_2, ..., a_N] \in \mathbb{R}^{N \times H \times W}$ describes whether the current pixel and adjacent pixels belong to the same instances or not, where $a_n (n = 1, 2, ...N)$ denotes the different adjacent relations between the current pixel and its n order adjacent pixel. Given the segmentation groundtruth $y \in \mathbb{R}^{H \times W}$, a pixel affinity $a_{n,i} \in a_n$ at the i^{th} pixel of the image I is formulated as

$$a_{n,i} = \begin{cases} 0, & \text{if } y_i \neq y_{i+n} \\ 1, & \text{if } y_i = y_{i+n}, \end{cases} \tag{1}$$

where y_i and y_{i+n} are the instance segmentation IDs of paired pixels i and $i + n$. 1 means that pixel i and $i + n$ belong to one instance, while 0 means the opposite.

Considering the segmentation groundtruth with sparse instance-level annotations, we regard the unannotated region (including the background) as an identical instance ID. The generated affinity label is divided into annotated region R^L_a and unannotated region R^U_a. If the pixel affinity $a_{n,i}$ is located in the annotated region R^L_a, at least one of the pixels i and $i + n$ is located in the annotation region R^L, *i.e.*,

$y_i \in R^L$ or $y_{i+n} \in R^L$. Therefore, the annotated affinity region R_a^L is slightly larger than the annotated segmentation region R^L.

4. Cross-Representation Affinity Consistency

In this section, we introduce the proposed sparsely supervised biomedical instance segmentation framework. The input of our framework can be either 2D images or 3D volumes. Here we illustrate it using a 2D image example for easy visualization. As shown in Fig. 2, our framework consists of two main parts. The first part contains two parallel networks to build affinity perturbation by predicting different representations: IAM and EAM (detailed in 4.1). The second part contains an affinity cross-supervision mechanism by a pseudo-label noise filtering scheme to select the highly confident region of the two affinity maps to supervise each other for consistency learning (detailed in 4.2).

4.1. Affinity Perturbation

Unlike the common perturbation consistency strategy, we build the perturbation between two affinity representations predicted by two parallel networks. Following the existing affinity learning methods [11, 31], we adopt a network $f(\theta_1)$ to directly predict an implicit affinity map[1] $\hat{A}^1 = [\hat{a}_1^1, \hat{a}_2^1, ..., \hat{a}_N^1]$. For accurate pixel affinities, the network $f(\theta_1)$ focuses on the structure information of instance boundaries.

In contrast to implicit learning, explicitly modeled affinity has been widely studied [22, 14] by calculating the pairwise relationships between pixel embeddings. Given the limited number of instances, we only impose local constraints on the adjacent instances instead of constraining all instances by a discriminative loss [9]. We adopt a network $f(\theta_2)$ to predict an embedding map $E \in \mathbb{R}^{D\times H\times W}$, where D is the number of channels of the last layer of the network. Each pixel i of the image I is mapped into a pixel embedding vector $e_i \in \mathbb{R}^D$, which is a D-dimensional feature representation. We then adopt a cosine distance to calculate the relationship between pixel embeddings and converted the embedding map into an explicit affinity map $\hat{A}^2 = [\hat{a}_1^2, \hat{a}_2^2, ..., \hat{a}_N^2]$. The transformation from a paired of pixel embeddings to a pixel affinity $\hat{a}_{n,i}^2$ is formulated as

$$\hat{a}_{n,i}^2 = \frac{e_i^T e_{i+n}}{\|e_i\|_2 \|e_{i+n}\|_2}. \tag{2}$$

The network $f(\theta_2)$ pays more attention to extracting the semantic instance information to discriminate pixels belonging to different instances in the feature space.

[1]We call it implicit affinity map to distinguish from the one explicitly calculated from the embedding map below.

We adopt the MSE loss to supervise both $f(\theta_1)$ and $f(\theta_2)$ in the annotated region R_a^L of the affinity label A:

$$\begin{aligned} L_s &= L_{s1} + L_{s2} \\ &= \frac{1}{|R_a^L|} \sum_{i\in R_a^L} \sum_{n=1}^{N} (\left\|\hat{a}_{n,i}^1 - a_{n,i}\right\|_2 + \left\|\hat{a}_{n,i}^2 - a_{n,i}\right\|_2), \end{aligned} \tag{3}$$

where L_{s1} and L_{s2} represent the loss functions for $f(\theta_1)$ and $f(\theta_2)$, respectively.

4.2. Affinity Cross-Supervision

In order to combine the advantages of the two kinds of affinity modeling, we propose an affinity cross-supervision mechanism to enforce the consistency between the IAM and the EAM. Since the two affinity maps may suffer from prediction errors during the initial training phase, we propose a pseudo-label noise filtering scheme by integrating two entropy-based decision strategies, *i.e.*, adaptive threshold and prediction by a CPSN, to select highly confident pixel affinities to avoid the collapse of the consistency learning. For a pixel affinity $a_{n,i}$, its entropy is computed by

$$h_{n,i} = -a_{n,i} \log (a_{n,i}) - (1 - a_{n,i}) \log (1 - a_{n,i}). \tag{4}$$

Adaptive threshold. As a heuristic strategy, we can determine whether the pixel affinity is confident by the following indicator:

$$\delta_{n,i}^{thres} = \begin{cases} 1, & \text{if } h_{n,i} < \gamma_t \\ 0, & \text{others}, \end{cases} \tag{5}$$

where γ_t refers to the entropy threshold at the t^{th} training iteration, which is the quantile of $h_{n,i}$ corresponding to α_t to limit unreliable pixels with top α_t entropy. 1 means the pixel affinity is confident with low entropy, while 0 means the opposite.

Since the confident pixels gradually increase during the training procedure, α_t is dynamically adjusted by a power function:

$$\alpha_t = \alpha_0 \left(1 - t/T\right)^p, \tag{6}$$

where α_0 is the initial quantile and is set to 5%, p is the power and is set as 1.5, t is the current training iteration, and T is total iterations.

Confidence pixel selection network. Inspired by approaches that address pseudo label errors using auxiliary networks [35, 21], we employ a binary-classification network called CPSN to identify highly confident pixels, which is trained only using the annotated regions of the groundtruth error map, as shown in Fig. 2 (d). Specifically, we sample five segmentation models under different training iterations of $f(\theta_1)$ and $f(\theta_2)$ respectively, and add a certain rate of dropout to these models. We then use the prediction from these models to simulate various segmentation errors. Using this kind of simulation data, we train

a robust CPSN to locate errors and generate highly confident pixels. We provide training details (including the loss L_{CPSN}) in the supplementary material.

Given the input image I, affinity map A and affinity entropy map H, the CPSN can predict a binary error map $B \in \mathbb{R}^{N\times H\times W}$ with the same size as A:

$$B = \mathrm{CPSN}\left(I \oplus A \oplus H\right), \tag{7}$$

where $\oplus$ is the concatenation operation.

We then determine whether the pixel affinity is confident by the following indicator:

$$\delta_{n,i}^{cpsn} = 1 - b_{n,i}, \tag{8}$$

where $b_{n,i} \in B$ is the pixel-level prediction of CPSN.

Pseudo-label noise filtering. We integrate the above two decision strategies to obtain more robust highly confident pixels, which is described by $\delta_{n,i} = \delta_{n,i}^{thres}\delta_{n,i}^{cpsn}$. After obtaining the confident pixels of the IAM (indicated by $\delta_{n,i}^{1}$) and the EAM (indicated by $\delta_{n,i}^{2}$), we utilize their confident pixels on the unannotated affinity region R_a^U to cross-supervise each other. The consistency loss is formulated as

$$\begin{aligned} L_c &= L_{c1} + L_{c2} \\ &= \frac{1}{|R_a^U|}\sum_{i\in R_a^U}\sum_{n=1}^{N}(\left\|\hat{a}_{n,i}^{1} - \hat{a}_{n,i}^{2}\right\|_2 \delta_{n,i}^{2} \\ &\qquad + \left\|\hat{a}_{n,i}^{2} - \hat{a}_{n,i}^{1}\right\|_2 \delta_{n,i}^{1}), \end{aligned} \tag{9}$$

where L_{c1} and L_{c2} represent the consistency loss functions for $f(\theta_1)$ and $f(\theta_2)$.

4.3. Overall Optimization

The overall objection function is the combination of losses on the annotated region and the unannotated region, which terms

$$L = L_s + \lambda L_c, \tag{10}$$

where λ is the trade-off weight and is set to make the two loss value ranges comparable.

Following [14], we adopt the network $f(\theta_2)$ to predict an explicit affinity map in the inference stage. The affinity map is converted into final instance segmentation results by different post-processing algorithms. The 3D results and 2D results are generated by Waterz [11] (50% quantile and 0.5 threshold) and Mutex [45] post-processing algorithms, respectively. Since affinities calculated by these two different normalization schemes (sigmoid vs cosine) belong to different intervals, we use a RELU function to map the output value into $[0, 1]$, for Waterz post-processing algorithm.

5. Experiments

5.1. Datasets and Metrics

We conduct evaluations on four representative biomedical datasets with corresponding commonly used metrics.

AC3. The AC3 dataset is a popular electron microscopy (EM) neuron dataset, imaged from the mouse somatosensory cortex at $3 \times 3 \times 29\ nm$ resolution. It is a subset of the Kasthuri [16] dataset, and contains 256 volumetric images (1024×1024). We adopt the top 100 sections as the labeled set and sparsely select different numbers of neuron instances as the annotations. At the same time, the bottom 50 sections are adopted for evaluation. We adopt two widely used metrics in the field of EM image segmentation to quantitatively evaluate the result, *i.e.*, Variation of Information (VOI) [34] and Adapted Rand Error ($ARAND$) [39]. The VOI is defined as a sum of another two metrics VOI_{split} and VOI_{merge} which indicate the split and merge errors, respectively. Note that smaller values of these metrics indicate better performance.

CREMI-C. The CREMI dataset is from the CREMI challenge [8] for neuron segmentation in EM volumes. It is composed of three volumetric datasets (CREMI-A, CREMI-B, and CREMI-C) imaged from the adult drosophila brain at $4 \times 4 \times 40\ nm$ resolution. Each of the three datasets contains 125 images.Given that the neurites in CREMI-A and CREMI-C are mostly homogeneous in morphology, we evaluate our method on the CREMI-C dataset, which contains more challenging neuron types. We adopt the top 75 sections as the labeled set and the bottom 50 sections for evaluation. We adopt the same metrics used for the AC3 to evaluate the results on the CREMI-C dataset.

CVPPP. The A1 sequence of the CVPPP [36] dataset contains leaves with complex shapes and severe occlusions, which is widely used to evaluate biomedical instance segmentation. We randomly sample 108 images for training and 20 images for testing, where each image is with a size of 530×500. We adopt the Symmetric Best Dice (SBD) and the absolute Difference in Counting ($|DiC|$) metrics [36] to evaluate the quantitative result.

BBBC039V1. The BBBC039V1 dataset [32] contains 200 fluorescence microscopy images with a resolution of 696×520 pixels. The dataset focus on the U2OS cells with various shape and density. Following the official data split, we use 100 images for training, 50 for validation, and the rest 50 for testing. We adopt four metrics for cell segmentation for quantitative evaluation, *i.e.*, Aggregated Jaccard Index (AJI) [39], object-level F1 score ($F1$) [4], Panoptic Quality (PQ) [19], and pixel-level Dice score ($Dice$).

5.2. Implementation Details

We adopt two different backbone networks, *i.e.*, 3D U-Net [23] and 2D Residual U-Net for the 3D volumetric datasets (AC3, CREMI-C) and 2D datasets (CVPPP, BBBC039V1), respectively. Both the network $f(\theta_1)$ and $f(\theta_2)$ are based on the same backbone network but different in the last convolution layer. We train these networks using Adam optimizer [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, a

	Methods	AC3				CREMI-C			
		$VOI_{Split}\downarrow$	$VOI_{Merge}\downarrow$	$VOI\downarrow$	$ARAND\downarrow$	$VOI_{Split}\downarrow$	$VOI_{Merge}\downarrow$	$VOI\downarrow$	$ARAND\downarrow$
Semi.	Vanilla	0.422	5.403	5.825	0.911	0.929	2.117	3.046	0.580
	PseudoL-hard [41]	0.089	5.955	6.044	0.920	0.863	1.462	2.325	0.311
	PseudoL-soft [50]	0.851	3.744	4.595	0.853	0.950	0.894	1.844	0.230
	MT [42]	0.863	2.058	2.921	0.558	0.916	1.027	1.943	0.211
	π-model [25]	0.847	1.867	2.714	0.562	0.985	0.918	1.903	0.227
	UA-MT [49]	0.720	1.860	2.580	0.426	0.978	0.863	1.841	0.227
	SASSNet [24]	1.160	1.180	2.340	0.324	0.935	0.814	1.749	0.193
	SSNS [13]	0.702	0.527	1.229	0.120	0.934	0.775	<u>1.709</u>	0.175
	SPOCO [46]	1.608	1.349	2.957	0.236	1.832	0.896	2.728	0.303
Sparse.	Vanilla	0.930	0.237	1.167	0.111	1.613	0.451	2.064	0.187
	MT* [42]	0.699	0.432	1.132	0.092	1.375	0.559	1.934	0.173
	UA-MT* [49]	0.785	0.251	1.036	0.082	1.502	0.333	1.835	0.155
	CCT* [37]	0.859	0.238	1.097	0.088	1.425	0.463	1.888	0.173
	CPS* [7]	0.718	0.248	<u>0.966</u>	<u>0.067</u>	1.384	0.398	1.782	<u>0.158</u>
	Ours	0.642	0.260	**0.903**	**0.064**	1.032	0.587	**1.620**	**0.153**

Table 1. Quantitative comparison of segmentation results on AC3 and CREMI-C datasets using 5 densely annotated sections for semi-supervision and randomly annotated instances for sparse instance-level supervision. The number of annotated voxels is the same for the two settings. Vanilla refers to training models only using available groundtruth. The top right corner '*' indicates these methods are originally proposed in the semi-supervision but adapted into the sparse instance-level supervision. Note that we adopt the same backbone ResUnet [23] in all methods for a fair comparison. The best results and the second-best results are highlighted in bold and underlined.

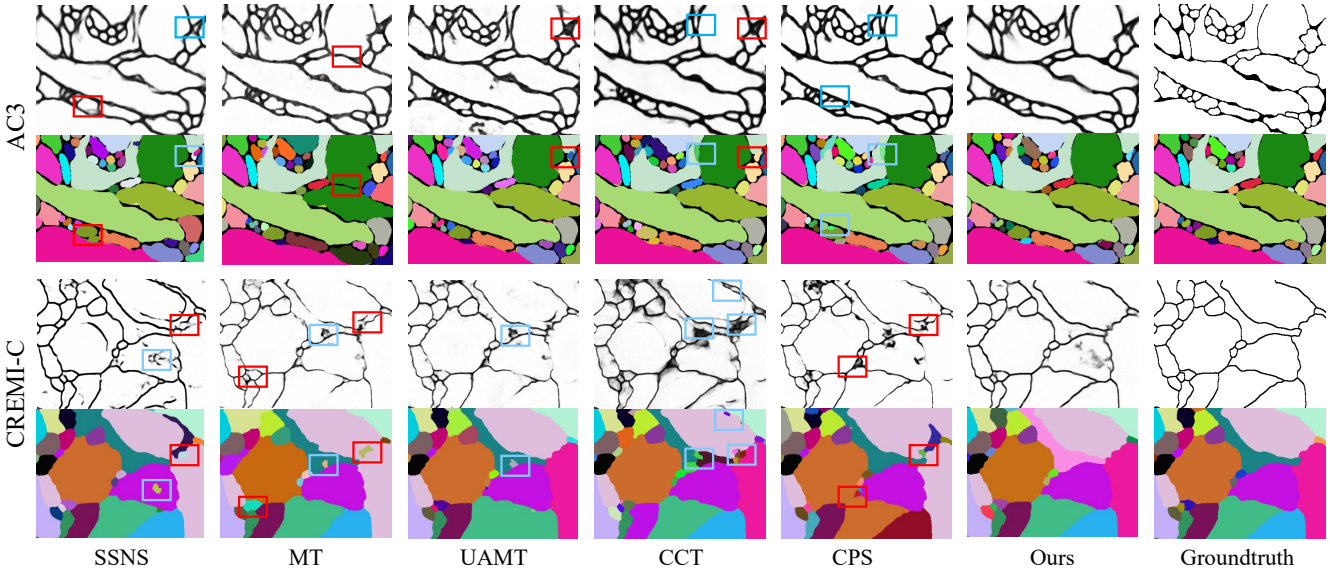

Figure 3. 2D Visual comparison of different methods on AC3 and CREMI-C datasets. The first row and the second row of each dataset are the affinity map and the corresponding instance segmentation result. Red and blue boxes indicate merge and split errors, respectively.

learning rate of $1e-4$, and a batch size of 2 on one NVIDIA TitanXP GPU for 200K iterations.

5.3. Comparison Methods

The evaluation of our method is performed against three kinds of baselines:

(1) For the 3D volumetric datasets, the sparse instance-level supervision and the common semi-supervision contain different 3D structure information of instances in the axial direction. For a comprehensive evaluation, we compare our proposed sparsely supervised method with a number of existing semi-supervised methods that are widely used for 3D datasets, including: PseudoL-hard [41], PseudoL-soft [50], mean-teacher (MT) [42], π-model [25], UA-MT [49], SASSNet [24] and SSNS [13].

(2) SPOCO [46] provides a baseline solution for the sparsely supervised segmentation task. We use the official code and configurations to reproduce the results on 2D datasets (CVPPP and BBBC039V1), and adapt this method to 3D datasets (AC3 and CREMI-C).

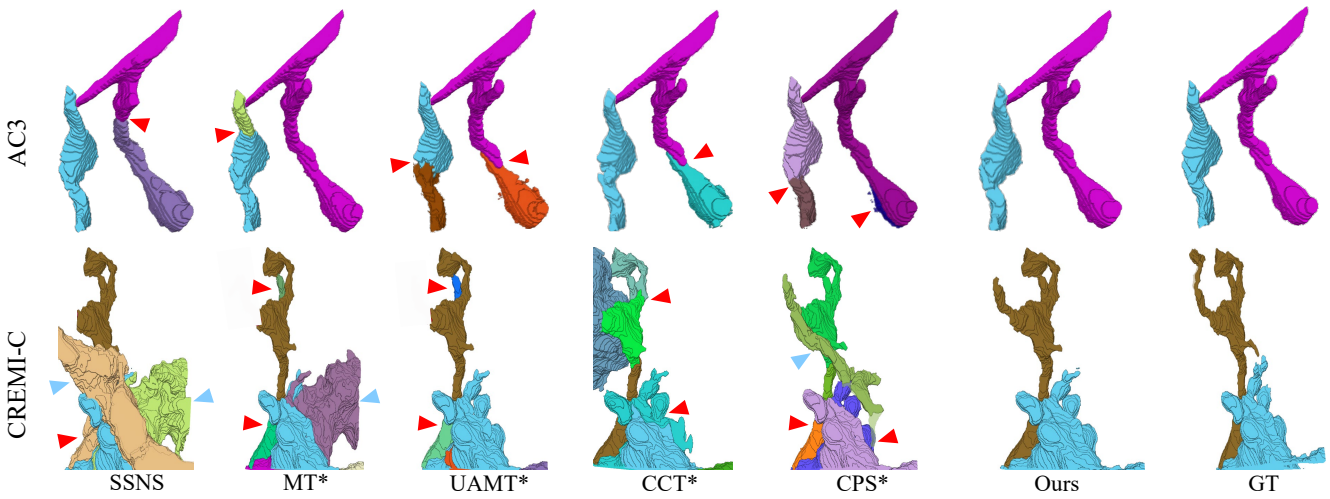

Figure 4. 3D Visual comparison of different methods on AC3 and CREMI-C datasets. Red and blue arrows indicate split and merge errors, respectively.

(3) We adapt a variety of advanced semi-supervised methods to the sparsely supervised segmentation task, including mean-teacher (MT) [42], uncertainty-aware mean teacher model(UA-MT) [49], cross-consistency training (CCT) [37] and cross pseudo supervision (CPS) [7].

5.4. Results on 3D Datasets

We list an extensive quantitative comparison on the AC3 and the CREMI-C datasets in Table 1. We conduct the semi-supervised experiments with 5 densely annotated sections, and the sparsely supervised experiments with 23 and 33 randomly annotated instances for AC3 and CREMI-C datasets respectively. The number of annotated voxels is the same for the two settings. From the results in Table 1, we can observe that:

(1) With the same number of annotated voxels, sparsely supervised methods perform better than corresponding semi-supervised methods by a large margin. Specifically, our results outperform the state-of-the-art semi-supervised results from SSNS [13] by 26.4% and 5.2% for the VOI metric on the AC3 and CREMI-C datasets.

(2) For the EM neuron segmentation datasets, the sparse instance annotations are not conducive to the training of SPOCO [46] based on metric learning. Therefore, SPOCO [46] does not perform well on these two datasets.

(3) Our method achieves better performance than other adapted methods in the sparse instance-level supervision setting. Specifically, We achieve 6.3% and 9.2% improvement from the key VOI metric on the AC3 and CREMI-C datasets, compared with the second-best method. It demonstrates the superiority of the proposed affinity consistency regularization.

We show 2D and 3D visual comparison results in Fig. 3 and Fig. 4, respectively. As can be seen, our proposed method predicts the affinity map with higher fidelity than other methods, which significantly reduces the split and merge errors. We also provide 3D visual comparison results in the supplementary material. As can be seen, our results contain fewer merge and split errors, and maintain the accuracy of neuron structures, compared with other methods.

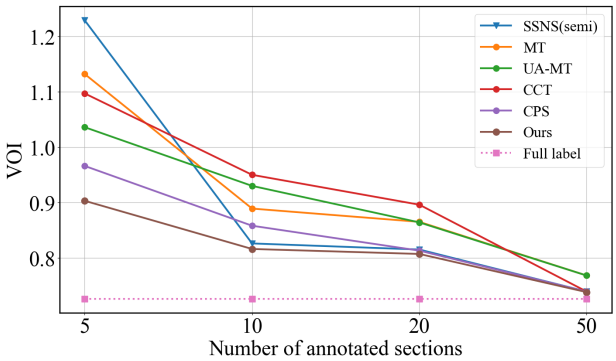

Figure 5. Comparison of different methods on the AC3 dataset under different annotation amounts converted into the number of sections for counting, where each section contains 1×10^6 pixels.

We further compare different methods on the AC3 dataset under different numbers of annotated voxels. We convert the number of voxels to the number of sections for counting, where each section contains about 1×10^6 voxels. The results are shown in Fig. 5. As more annotations are available, the performance gap between different methods gradually narrows, but the superiority of our method can still be observed. Especially, the state-of-the-art semi-supervised method SSNS performs worse than sparsely supervised methods under very limited annotations, due to the

Methods	10% instances		40% instances	
	SBD ↑	|DiC| ↓	SBD ↑	|DiC| ↓
CPS [7] (semi)	83.2	1.8	**87.6**	1.1
SPOCO [46]	70.6	3.0	82.1	2.0
Vanilla	69.8	6.0	78.0	1.9
MT* [42]	78.5	1.4	84.3	1.3
UA-MT* [49]	79.6	1.4	84.5	1.1
CCT* [37]	78.8	1.9	83.9	1.6
CPS* [7]	80.7	1.8	85.5	1.1
Ours	**83.6**	**1.1**	86.6	**0.8**

Table 2. Quantitative comparison results on the CVPPP dataset. Results in the first row are based on the semi-supervision setting.

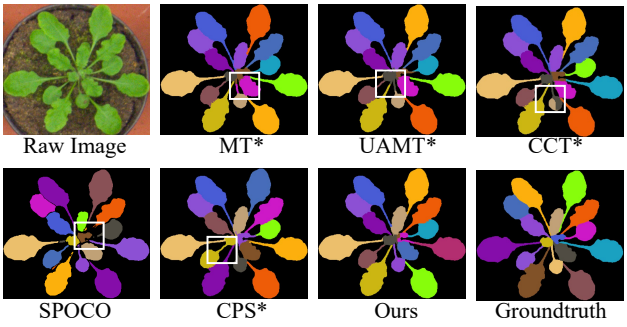

Figure 6. Visual comparison of different methods on the CVPPP dataset with 40% instances. The segmentation errors are highlighted by white boxes.

incomplete 3D structure information of instances.

5.5. Results on 2D Datasets

We further validate our method on two 2D datasets, *i.e.,* CVPPP and BBBC039V1, and their quantitative results are shown in Table 2 and Table 3, respectively. Compared with other adapted methods in the sparse instance-level supervision setting, our method achieves the best performance on the CVPPP dataset with 10% and 40% annotated instances, and the BBBC039V1 dataset with 1% and 10% annotated instances. Visual results on these two datasets also illustrate the superior performance of our method with fewer segmentation errors, as shown in Fig. 6 and Fig. 7.

Although the sparse instance-level supervision and the common semi-supervision have no difference in the instance structure for 2D datasets, we also evaluate the semi-supervised setting (with the same annotated pixels) on these two datasets. Our proposed method performs better than the state-of-the-art semi-supervised method CPS [7], when the annotation amount is very limited. Meanwhile, our method significantly outperforms the existing sparely supervised method SPOCO in all experiments. As more annotation amount is available, the semi-supervised method has a better performance. Therefore, the sparse instance-level supervision is still a meaningful complement to the common semi-supervision when instances are so crowded that dense annotation is laborious and the amount of annotations is very limited, for the 2D datasets.

	Methods	AJI ↑	Dice ↑	F1 ↑	PQ ↑
1% instances	CPS [7] (semi)	0.8120	0.9033	**0.9595**	0.8168
	SPOCO [46]	0.6896	0.8478	0.8492	0.6933
	Vanilla	0.5631	0.6411	0.7570	0.4890
	MT* [42]	0.7716	0.8260	0.9018	0.7533
	UA-MT* [49]	0.8065	0.8738	0.9439	0.7928
	CCT* [37]	0.8113	0.9136	0.9295	0.7830
	CPS* [7]	**0.8323**	0.9185	0.9444	0.8153
	Ours	0.8303	**0.9338**	0.9442	**0.8222**
10% instances	CPS [7] (semi)	0.8393	0.9189	**0.9598**	**0.8437**
	SPOCO [46]	0.7488	0.8746	0.9263	0.7503
	Vanilla	0.5985	0.7177	0.7709	0.5652
	MT* [42]	0.8416	0.9167	0.9338	0.8283
	UA-MT* [49]	0.8404	0.9118	0.9380	0.8314
	CCT* [37]	0.8309	0.9279	0.9327	0.7975
	CPS* [7]	0.8463	0.9366	0.9408	0.8174
	Ours	**0.8545**	**0.9469**	0.9450	0.8343

Table 3. Quantitative comparison on the BBBC039V1 dataset.

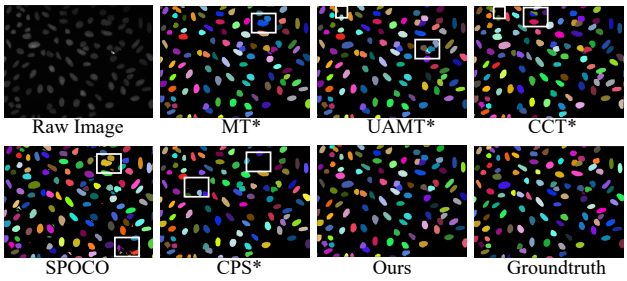

Figure 7. Visual comparison of different methods on the BBBC039V1 dataset with 10% instances.

5.6. Ablation Studies

Method	1	5	10	20	50	100
Vanilla	2.289	1.167	0.982	0.871	0.779	0.726
Ours	1.271	0.903	0.816	0.807	0.738	0.717

Table 4. Ablation study on the annotation amount which is converted into the number of sections for counting, where each section contains about 1×10^6 voxels. The key metric VOI is used to evaluate the results.

We conduct ablation studies on the impacts of each component of the proposed framework on the AC3 dataset.

Annotation amounts. We conduct an ablation study to investigate the effects of different amounts of annotated instances. As shown in Table 4, our proposed method consistently outperforms the vanilla model trained only using available annotated instances. When the annotation amount is small (*e.g.,* the number of annotated voxels is about a sec-

$f(\theta_1)$	$f(\theta_2)$	cross-supervision		$VOI\downarrow$	$ARAND\downarrow$
		Threshold	CPSN		
IAM	IAM	✓	✓	0.926	0.074
EAM	EAM	✓	✓	0.927	0.068
IAM	EAM	✗	✗	0.916	0.068
IAM	EAM	✓	✗	0.906	0.068
IAM	EAM	✗	✓	0.909	0.067
IAM	EAM	✓	✓	**0.903**	**0.065**

Table 5. Ablation study on different components of the proposed affinity consistency regularization. IAM and EAM represent the different kinds of affinity modeling for the two networks $f(\theta_1)$ and $f(\theta_2)$. 'Threshold' and 'CPSN' denote the two different strategies to select highly confident regions in cross-supervision.

Method	$VOI_{Split}\downarrow$	$VOI_{Merge}\downarrow$	$VOI\downarrow$	$ARAND\downarrow$
Ours-$f(\theta_1)$	0.709	0.198	0.907	0.071
Ours-$f(\theta_2)$	0.687	0.218	**0.903**	**0.065**

Table 6. Ablation study on which affinity map used for inference.

α_0	0.01	0.05	0.1
VOI↓ / ARAND↓	0.933 / 0.075	**0.903 / 0.065**	0.924 / 0.069
Dropout rate	0	0.1	0.3
VOI↓ / ARAND↓	0.920 / 0.066	**0.903 / 0.065**	0.923 / 0.067
λ	0.01	0.1	1
VOI↓ / ARAND↓	0.914 / 0.066	**0.903 / 0.065**	1.258 / 0.106

Table 7. Ablation study on key parameters of the proposed framework. VOI/ARAND are adopted as metrics.

tion), our method significantly improves the performance of the vanilla model. The performance gap narrows as more annotations are available.

Affinity consistency regularization. As shown in Table 5, we conduct an ablation study on the proposed affinity consistency learning scheme. We compare different combinations of affinity modeling for the two networks $f(\theta_1)$ and $f(\theta_2)$ by directly enforcing their consistency. It can be observed that the combination of IAM and EAM achieves the best performance. Furthermore, we demonstrate the effectiveness of pseudo-label noise filtering with different decision strategies. Both strategies have been shown to improve performance, and their integration can obtain more robust highly confident regions and achieves the best performance. Moreover, We have compared our method with an equivalent implementation of [14] in Table 5, *i.e.*, two network branches both use EAM. Our method achieves better performance.

Affinity map used in the inference stage. The two affinity maps predicted by $f(\theta_1)$ and $f(\theta_2)$ have different performances, as shown in Table 6. The EAM predicted by $f(\theta_2)$ performs slightly better, so we adopt it in the inference stage.

Ablation study on hyper-parameters. We conducted ablation experiments to evaluate the impact of key hyper-parameters, of the proposed framework, as shown in Table 7. For the proposed pseudo-label noise filtering scheme, we test different values of hyper-parameters α_0 for adaptive threshold and different dropout rates used in the CPSN training. Also, we provide the ablation study on different loss weights λ.

6. Conclusion

In this paper, we propose an effective sparsely supervised biomedical instance segmentation framework via learning cross-representation affinity consistency. The proposed framework builds the perturbation consistency between an implicit affinity map and an explicit affinity map with an affinity cross-supervision mechanism. We conduct extensive experiments on 3D volumetric datasets and 2D datasets to demonstrate the superiority of our proposed framework over the existing sparsely supervised method. Meanwhile, for the first time, we validate that sparsely supervised methods can better utilize the 3D structure information of instances and perform better than semi-supervised methods for volumetric datasets.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62021001.

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. 2014. 3

[2] Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature Methods*, 14(2):101–102, 2017. 3

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019. 2

[4] Hao Chen, Xiaojuan Qi, Lequan Yu, Qi Dou, Jing Qin, and Pheng-Ann Heng. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis*, 36:135–146, 2017. 5

[5] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *CVPR*, 2016. 2

[6] Qi Chen, Mingxing Li, Jiacheng Li, Bo Hu, and Zhiwei Xiong. Mask rearranging data augmentation for 3d mitochondria segmentation. In *MICCAI*, 2022. 1

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, 2021. 1, 2, 3, 6, 7, 8

[8] CREMI. Miccai challenge on circuit reconstruction from electron microscopy images. https://cremi.org/, 2016. 5

[9] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. In *CVPR*, 2017. 2, 4

[10] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. In *BMVC*, 2020. 3

[11] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. 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, 2019. 2, 3, 4, 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[13] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022. 1, 2, 6, 7

[14] Wei Huang, Shiyu Deng, Chang Chen, Xueyang Fu, and Zhiwei Xiong. Learning to model pixel-embedded affinity for homogeneous instance segmentation. In *AAAI*, 2022. 3, 4, 5, 9

[15] Wei Huang, Xiaoyu Liu, Zhen Cheng, Yueyi Zhang, and Zhiwei Xiong. Domain adaptive mitochondria segmentation via enforcing inter-section consistency. In *MICCAI*, 2022. 1

[16] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015. 5

[17] Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson WH Lau. Dual student: Breaking the limits of the teacher in semi-supervised learning. In *ICCV*, 2019. 3

[18] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[19] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 5

[20] Victor Kulikov and Victor Lempitsky. Instance segmentation of biological images using harmonic embeddings. In *CVPR*, 2020. 1, 2

[21] Donghyeon Kwon and Suha Kwak. Semi-supervised semantic segmentation with error localization network. In *CVPR*, 2022. 4

[22] Kisuk Lee, Ran Lu, Kyle Luther, and H Sebastian Seung. Learning and segmenting dense voxel embeddings for 3d neuron reconstruction. *IEEE Transactions on Medical Imaging*, 40(12):3801–3811, 2021. 1, 2, 4

[23] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017. 5, 6

[24] Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *MICCAI*, 2020. 6

[25] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020. 3, 6

[26] 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. 1, 2

[27] Xiaoyu Liu, Bo Hu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Efficient biomedical instance segmentation via knowledge distillation. In *MICCAI*, 2022. 2

[28] Xiaoyu Liu, Bo Hu, Mingxing Li, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. A soma segmentation benchmark in full adult fly brain. In *CVPR*, 2023. 2

[29] Xiaoyu Liu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Biological instance segmentation with a superpixel-guided graph. In *IJCAI*, 2022. 2

[30] Xiaoyu Liu, Yueyi Zhang, Zhiwei Xiong, Chang Chen, Wei Huang, Xuejin Chen, and Feng Wu. Learning neuron stitching for connectomics. In *MICCAI*, 2021. 1

[31] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *ECCV*, 2018. 2, 3, 4

[32] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 1, 5

[33] Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, and Feng Wu. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *CVPR*, 2023. 2

[34] Marina Meilă. Comparing clusterings by the variation of information. In *LTKM*. 2003. 5

[35] Robert Mendel, Luis Antonio De Souza, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *ECCV*, 2020. 4

[36] 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. 1, 5

[37] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[38] Christian Payer, Darko Štern, Thomas Neff, Horst Bischof, and Martin Urschler. Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks. In *MICCAI*, 2018. 2

[39] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. 5

[40] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. 2016. 3

[41] Eichi Takaya, Yusuke Takeichi, Mamiko Ozaki, and Satoshi Kurihara. Sequential semi-supervised segmentation for serial electron microscopy image with small number of labels. *Journal of Neuroscience Methods*, 351:109066, 2021. 6

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. 1, 2, 3, 6, 7, 8

[43] Srinivas C. Turaga, Kevin L. Briggman, Moritz Helmstaedter, Winfried Denk, and H. Sebastian Seung. Maximin affinity learning of image segmentation. In *NeurIPS*, 2009. 2

[44] Srinivas C Turaga, Joseph F Murray, Viren Jain, Fabian Roth, Moritz Helmstaedter, Kevin Briggman, Winfried Denk, and H Sebastian Seung. Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural computation*, 22(2):511–538, 2010. 2

[45] Steffen Wolf, Alberto Bailoni, Constantin Pape, Nasim Rahaman, Anna Kreshuk, Ullrich Köthe, and Fred A Hamprecht. The mutex watershed and its objective: Efficient, parameter-free graph partitioning. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3724–3738, 2020. 5

[46] Adrian Wolny, Qin Yu, Constantin Pape, and Anna Kreshuk. Sparse object-level supervision for instance segmentation with pixel embeddings. In *CVPR*, 2022. 2, 6, 7, 8

[47] Jingru Yi, Hui Tang, Pengxiang Wu, Bo Liu, Daniel J Hoeppner, Dimitris N Metaxas, Lianyi Han, and Wei Fan. Object-guided instance segmentation for biological images. In *AAAI*, 2020. 2

[48] Jingru Yi, Pengxiang Wu, Hui Tang, Bo Liu, Qiaoying Huang, Hui Qu, Lianyi Han, Wei Fan, Daniel J Hoeppner, and Dimitris N Metaxas. Object-guided instance segmentation with auxiliary feature refinement for biological images. *IEEE Transactions on Medical Imaging*, 2021. 1

[49] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *MICCAI*, 2019. 1, 2, 3, 6, 7, 8

[50] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 6