DualRel: Semi-Supervised Mitochondria Segmentation from A Prototype Perspective

Huayu Mai[1*] Rui Sun[1*] Tianzhu Zhang[1,2,3†] Zhiwei Xiong[1,2] Feng Wu[1,2]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Deep Space Exploration Lab
{mai556, issunrui}@mail.ustc.edu.cn, {tzzhang, zwxiong, fengwu}@ustc.edu.cn

Abstract

Automatic mitochondria segmentation enjoys great popularity with the development of deep learning. However, existing methods rely heavily on the labor-intensive manual gathering by experienced domain experts. And naively applying semi-supervised segmentation methods in the natural image field to mitigate the labeling cost is undesirable. In this work, we analyze the gap between mitochondrial images and natural images and rethink how to achieve effective semi-supervised mitochondria segmentation, from the perspective of reliable prototype-level supervision. We propose a novel end-to-end dual-reliable (DualRel) network, including a reliable pixel aggregation module and a reliable prototype selection module. The proposed DualRel enjoys several merits. First, to learn the prototypes well without any explicit supervision, we carefully design the referential correlation to rectify the direct pairwise correlation. Second, the reliable prototype selection module is responsible for further evaluating the reliability of prototypes in constructing prototype-level consistency regularization. Extensive experimental results on three challenging benchmarks demonstrate that our method performs favorably against state-of-the-art semi-supervised segmentation methods. Importantly, with extremely few samples used for training, DualRel is also on par with current state-of-the-art fully supervised methods.

1. Introduction

Mitochondria, as one of the crucial organelles, are the primary energy providers for cell activities and are essential for metabolism. Quantification of mitochondrial morphology can not only promote basic scientific research (*e.g.*, cellular physiology [1, 5]), but also provide new insight for clinical diagnosis (*e.g.*, neurodegenerative diseases [20] and diabetes [24]). Recently, with the development of deep learning, semantic segmentation [2, 14, 18, 27, 30, 33] enables in-depth exploration of mitochondrial morphology from high-resolution electron microscopy (EM) images and make conspicuous achievements. However, their flexibility and scalability are limited in the actual deployment because the numerous cluttered irrelevant organelles that require labor-intensive manual discrimination and gathering by experienced domain experts [10, 21]. Therefore, we begin to turn attention to semi-supervised segmentation with the assumption that enormous unlabeled data is accessible, aiming to alleviate the data-hungry issue.

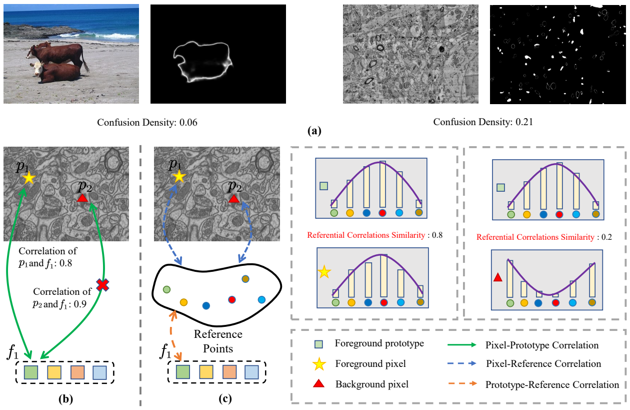

Confusion Density: 0.06
(a)
Confusion Density: 0.21
p_1
p_2
Correlation of p_1 and f_1: 0.8
Correlation of p_2 and f_1: 0.9
f_1
(b)
p_1
p_2
Reference Points
f_1
(c)
Referential Correlations Similarity : 0.8
Referential Correlations Similarity : 0.2
Foreground prototype
Foreground pixel
Background pixel
Pixel-Prototype Correlation
Pixel-Reference Correlation
Prototype-Reference Correlation

Figure 1. Illustration of our motivation. (a) shows the confusion map and density (*i.e.*, the expected inverse confidence per pixel) of mitochondrial and natural images. (b) shows the unreliability caused by direct pairwise prototype-pixel correlation that is conditioned only on visual similarity. (c) shows how to construct pixel-reference correlation to rectify the direct pairwise correlation in a referential correlation manner.

Semi-supervised segmentation enjoys great popularity in the field of natural images, and representative works such as CPS [3], which imposes pixel-level consistency regularization and establishes state-of-the-art performance. It naturally comes into mind to directly apply a CPS-like method to semi-supervised mitochondria segmentation. However, there exist a large gap between mitochondrial and natural images. As shown in Fig. 1 (a), we observe that the confusion density (*i.e.*, the expected inverse confidence per pixel) in mitochondrial images significantly surpasses counterpart in natural images, implying that directly employing

[*] Equal contribution
[†] Corresponding author

pixel-level consistency regularization as supervision signals on mitochondrial images will inevitably increase the risk of unreliability. The most intuitive example is that there exist considerable boundary regions in mitochondrial images, and the segmentation network is naturally equivocal for these regions, as proven in [15]. In this case, some relatively small mitochondria are easily overwhelmed by this ambiguity, leading to sub-optimal results.

In order to seek more reliable supervision signals to alleviate the undependable problem caused by pixel-level supervision, we draw inspiration from the inbuilt resistance to noisy pixels of prototypes and construct more robust and reliable *prototype-level supervision*. To achieve this goal, two issues need to be considered. (1) Unreliable pixels. Considering the cluttered background caused by under/overexposure and out-of-focus problems during EM imaging, the prototype inevitably absorbs unreliable pixels (*i.e.*, heterogeneous semantic clues) during the interaction of corresponding pixels with a suitable pattern. We argue that directly forcing pairwise prototype-pixel correlation is primarily at blame. As shown in Fig. 1 (b), due to the foreground-background ambiguity, the *foreground* prototype f_1 is erroneously closer to p_2 located in the *background* than counterpart point p_1 with similar pattern situated in the *foreground*. Therefore, it is highly desirable to suppress the unreliable pixels caused by the direct pairwise prototype-pixel correlation that is only conditioned on visual similarity during prototype learning process. (2) Unreliable prototypes. Intuitively, not all prototypes are equivalent for building prototype-level consistency regularization. For example, for a prototype that focuses on mitochondrial boundary patterns, the inherent unreliability of the pixels belonging to these patterns, as discussed above, will also taint the purity of this prototype with equivocality. Therefore, the prototype-level supervision signals should be further optimized to guarantee that the true reliable prototypes enjoy higher weights.

To mitigate the above issues, we rethink how to achieve effective consistency regularization for semi-supervised mitochondria segmentation, from the perspective of reliable prototype-level supervision. We propose a **Dual-Rel**iable (DualRel) network including a reliable pixel aggregation module and a reliable prototype selection module. **In the reliable pixel aggregation module** (RPiA), to learn the prototypes well without any explicit supervision, we carefully design the *referential correlation* to rectify the direct pairwise correlation, enabling the prototype absorb counterpart reliable pixels with the same semantic pattern during the interaction with the pixels. The main idea is, for each pixel/prototype, we can obtain the referential correlation (*i.e.*, a likehood vector) by comparing this pixel/prototype with a set of reliable reference points. In essence, the referential correlation reflects the consensus among reliable reference points with a broader receptive field and thus it encodes the relative semantic comparability of the reference points that can be relied upon, which is from a different perspective than the absolute pairwise prototype-pixel correlation. Intuitively, each pair of true prototype-pixel correlation (*e.g.*, the f_1-p_1 pair in Fig. 1 (c) derived from the prototypes and mitochondria images should be not only visually similar to each other (*i.e.*, high direct pairwise prototype-pixel correlation), but also similar to any other reference point (*i.e.*, similar referential correlation pair). Moreover, we assemble referential correlation into the cross-attention mechanism with the ability to capture long-range dependencies. In this case, the relatively equivocal pixels (*e.g.*, the f_1-p_2 pair in Fig. 1 (c) will be suppressed while the reliable ones are highlighted to reduce the correspondence noise. **In the reliable prototype selection module** (RPrS), in order to further evaluate the reliability of prototypes in constructing prototype-level consistency regularization, we draw inspiration from bayesian deep learning [12] and devise a reliability-aware consistency loss to pursue implicitly learn the reliability about each prototype in a data-driven way. In this way, the equivocal prototypes will be suppressed while the reliable ones are highlighted in the supervision signals.

In this work, our contributions can be concluded as follows: (1) To the best of our knowledge, this is the first work to rethink how to achieve effective consistency regularization for semi-supervised mitochondria segmentation, from the perspective of reliable prototype-level supervision. We analyze the gap between mitochondrial images and natural images, hoping our work will provide some insight for researchers in this field. (2) We propose a dual-reliable (DualRel) network in a unified framework. Specifically, we design the reliable pixel aggregation module to rectify the direct pairwise correlation, the reliable prototype selection module to further evaluate the reliability of prototypes in constructing prototype-level consistency regularization. (3) Extensive experimental results on three challenging benchmarks demonstrate that our method performs favorably against state-of-the-art semi-supervised segmentation methods. Importantly, with extremely few samples used for training, DualRel is also on par with current state-of-the-art fully supervised methods.

2. Related Work

2.1. Mitochondria Segmentation

Segmenting mitochondria in EM images is vital for researchers to explore cellular functions and subcellular activities. Rather than traditional methods utilizing hand-crafted features [14, 18, 27], DL-based networks have shown significant performance improvement on mitochondria segmentation. The pioneering work Lucchi et al. [19] design a deep neural network based on supervoxels to model shapes of mitochondria. Oztel et al. [26] propose a deep convolutional neural network with a sliding window strategy and post-processing steps to boost the performance. Wei et al. [35] introduce a 3D U-Net based network and release a new challenging dataset containing mitochondria images with higher resolution. However, all previous methods only focus on the fully-supervised training, which require a large amount of labeled data. In this work, we introduce semi-supervised learning into mitochondria segmentation, aiming to alleviate the data-hungry issue.

2.2. Semi-supervised Semantic Segmentation

With the development of semi-supervised learning, various kinds of semi-supervised semantic segmentation algorithms have been proposed. GAN-based methods [9,22,29] try to synthesize additional training data or generate the pseudo labels for unlabeled data by using adversarial loss. Recently, the consistency regularization [6,23,34] has been intensely studied in semi-supervised semantic segmentation. The main idea of this type of approach is that the intermediate features or predictions should maintain consistency across different semantic-preserving transformations on the input or different network initialization [36,37]. MT [31] is a representative work that first adopts the teacher network to guide the learning of the student network in semi-supervised semantic segmentation. Alternatively, CCT [25] introduces dual independent models to supervise each other with the soft probability output, while CPS [3] is supervised with hard pseudo segmentation maps. GCT [11] employs two segmentation networks with the same structure but different weight initialization and enforces the consistency between the predictions. However, the above pixel-level supervision-based methods perform poorly on semi-supervised mitochondria tasks.

3. Method

In this section, we first formulate the semi-supervised mitochondria segmentation task and present the overview of the proposed DualRel. Then we describe the details of the reliable pixel aggregation module (RPiA) and reliable prototype selection module (RPrs) of DualRel. Finally, the training and inference procedure are discussed.

3.1. Overview

In semi-supervised mitochondria segmentation task, we wish to train a segmentation network with training data $D^L \cup D^U$. In the labeled set $D^L = \{\mathbf{I}_i^L, \mathbf{Y}_i\}_{i=0}^N$, each image is associated with a ground truth label $\mathbf{Y} \in \{0,1\}^{H\times W}$, where 1 denotes foreground while 0 denotes background. In the unlabeled set $D^U = \{\mathbf{I}_j^U\}_{j=0}^M$, M images is provided without ground truth. Given an EM image $\mathbf{I}$ (for brevity, we omit the superscript L/U and subscript i/j), let $\mathbf{X} \in \mathbb{R}^{h\times w\times D}$ denotes the feature map extracted from feature extractor (*e.g.*, ResNet50 [8]), where h, w and D denote the height, width and channel number of the feature map, respectively. Subsequently, the feature map $\mathbf{X}$ is fed into the upsampling module (*e.g.*, U-Net [28]) which outputs pixel embedding $\mathbf{E} \in \mathbb{R}^{H\times W\times C}$ with the same spatial scale as original input image. As illustrated in Fig. 2, the proposed DualRel includes two branches of feature extractor and upsampling module, which are of the same structure but different weight initialization. Such a two-branch structure is devised to construct consistency supervision signals for unlabeled data, which is a popular paradigm in semi-supervised learning. Besides, a reliable pixel aggregation module and a reliable prototype selection module are introduced to construct prototype-level consistency regularization for semi-supervised mitochondria segmentation.

3.2. Reliable Pixel Aggregation

To construct prototype-level supervision, we prepend a set of learnable embeddings $\mathbf{F} = \{\mathbf{f}_k\}_{k=1}^K$ (referred to as mito filters). Each filter $\mathbf{f}_k$ is represented as a C-dimension vector to interact with the feature map $\mathbf{X}$ and absorb counterpart reliable pixels with the same semantic pattern. Then, we adopt a cross-attention mechanism [32] to realize the interaction and obtain mito features $\widetilde{\mathbf{F}} = \{\widetilde{\mathbf{f}}_k\}_{k=1}^K$.

Since the cross-attention requires a 1D sequence as the input, we first utilize a 1×1 convolution kernel to reduce the channel number of the feature map $\mathbf{X}$ from D to C, and then flatten the spatial dimensions to produce the feature sequence $\widetilde{\mathbf{X}} = \{\widetilde{\mathbf{x}}_i\}_{i=1}^{hw} \in \mathbb{R}^{hw\times C}$. Specifically, we denote mito filters $\{\mathbf{f}_k\}_{k=1}^K$ as queries, the feature map $\widetilde{\mathbf{X}}$ as keys and values. Formally,

$$\mathbf{q}_k = \mathbf{f}_k \mathbf{W}^Q, \mathbf{k}_i = \widetilde{\mathbf{x}}_i \mathbf{W}^K, \mathbf{v}_i = \widetilde{\mathbf{x}}_i \mathbf{W}^V, \tag{1}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{C\times C}$ are linear projections, $k = 1, 2, ..., K$ and $i = 1, 2, ..., hw$. Then the pairwise correlation $s_{k,i}$ between each query $\mathbf{q}_k$ and the i^{th} key $\mathbf{k}_i$ is given as

$$s_{k,i} = \frac{\exp(\beta_{k,i})}{\sum_{j=1}^{hw} \exp(\beta_{k,j})}, \beta_{k,i} = \frac{\mathbf{q}_k \mathbf{k}_i^\mathsf{T}}{\sqrt{C}}, \tag{2}$$

where $\sqrt{C}$ is a scaling factor to stabilize training and T refers to the transpose operation.

Due to the susceptibility of pairwise filter-pixel correlation, the mito filter inevitably absorbs unreliable pixels during the interaction. We carefully design the referential correlation assembled in cross-attention to mitigate such issue. Specifically, we define N reference points $\widetilde{\mathbf{X}}^R = \{\widetilde{\mathbf{x}}_n^R\}_{n=1}^N \in \mathbb{R}^{N\times C}$ (detailed in the supplementary material). Respectively calculating the filter-reference correlation and the pixel-reference correlation as same as Eq. (2):

$$\mathbf{S}_k^{FR} = \text{softmax}(\frac{(\mathbf{f}_k \mathbf{W}^Q)(\widetilde{\mathbf{X}}^R \mathbf{W}^K)^\mathsf{T}}{\sqrt{C}}), \tag{3}$$

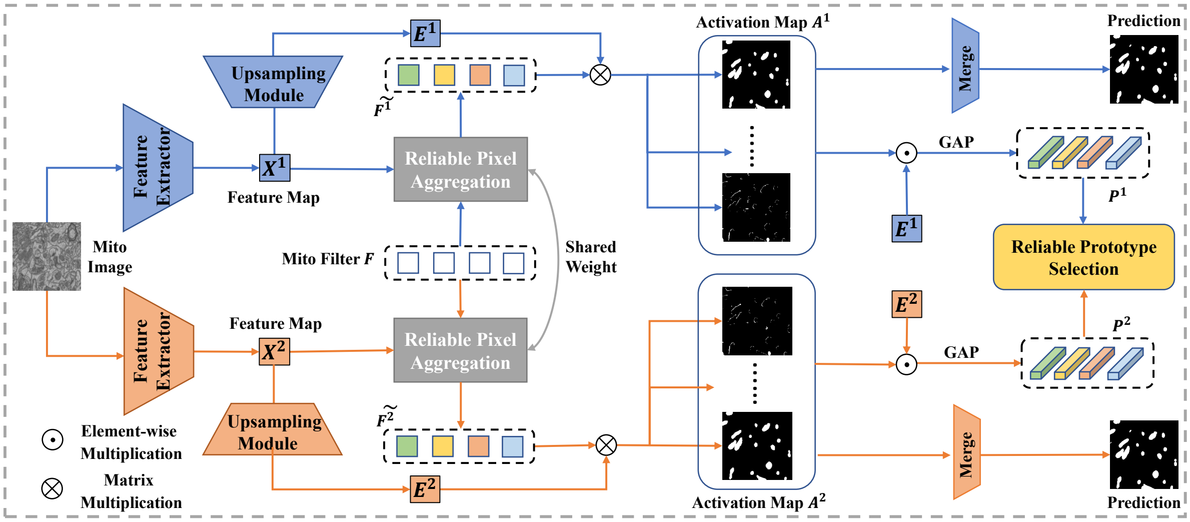


Figure 2. Illustration of the proposed DualRel. DualRel is mainly composed of a reliable pixel aggregation module to rectify the direct pairwise correlation, and a reliable prototype selection module responsible for further evaluating the reliability of prototypes in constructing prototype-level consistency regularization. "GAP" represents a global average pooling layer [16] and "Merge" is described in Sec. 3.4.

$$\mathbf{S}_i^{PR} = \mathrm{softmax}(\frac{(\widetilde{\mathbf{x}}_i \mathbf{W}^Q)(\widetilde{\mathbf{X}}^R \mathbf{W}^K)^\intercal}{\sqrt{C}}), \tag{4}$$

where $\mathbf{S}_k^{FR} \in \mathbb{R}^{1\times N}$ and $\mathbf{S}_i^{PR} \in \mathbb{R}^{1\times N}$ are regarded as the referential correlation. And then, we get the similarity between the referential correlations by:

$$c_{k,i} = \mathbf{S}_k^{FR} (\mathbf{S}_i^{PR})^\intercal, \tag{5}$$

which is used to rectify the direct pairwise correlation.

Then, the mito features $\{\widetilde{\mathbf{f}}_k\}_{k=1}^K$ can be got by blending values with the rectified correlations:

$$\widetilde{\mathbf{f}}_k = \sum_i^{hw} c_{k,i} \cdot s_{k,i} \cdot \mathbf{v}_i, \tag{6}$$

3.3. Reliable Prototype Selection

Before describing the reliable prototype selection mechanism, we first formulate the process of obtaining the mito prototypes $\mathbf{P} = \{\mathbf{p}_k\}_{k=1}^K$. First, calculating the activation maps $\mathbf{A} = \{\mathbf{A}_k\}_{k=1}^K \in \mathbb{R}^{K\times H\times W}$ by:

$$\mathbf{A}_k = \mathrm{sigmoid}(\mathbf{f}_k \mathbf{E}^\intercal), \tag{7}$$

where $\mathbf{A}_k$ denotes the activation map of the k^{th} mito feature on pixel embedding $\mathbf{E}$. Then, based on the activation maps $\mathbf{A}$, we generate mito prototypes $\{\mathbf{p}_k\}_{k=1}^K$ by global average pooling (GAP):

$$\mathbf{p}_k = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{A}_{k,i,j} \cdot \mathbf{E}_{i,j}. \tag{8}$$

Intuitively, not all prototypes are equivalent for building prototype-level consistency regularization. In order to further evaluate the reliability of prototypes in constructing prototype-level consistency regularization, we devise a reliability-aware consistency loss to pursue implicitly learn the reliability about each prototype in a data-driven way. Firstly, we estimate the reliability $\mathbf{R} = \{r_k\}_{k=1}^K$ for each prototype. In specific, we concatenate the mito features $\widetilde{\mathbf{f}}_k^1$ and $\widetilde{\mathbf{f}}_k^2$ (the superscript denotes which baranch the mito features from) in channel dimension and feed them into a two layers MLP followed by the sigmoid function, as in Eq. (9):

$$r_k = \mathrm{sigmoid}(\mathrm{MLP}(\mathrm{concat}(\widetilde{\mathbf{f}}_k^1, \widetilde{\mathbf{f}}_k^2))). \tag{9}$$

In order to guarantee that the true reliable prototypes enjoy higher weights and the equivocal ones are suppressed, we regard the $\mathbf{R}$ as a attenuation factor and get the reliable prototype consistency loss as:

$$\mathcal{L}_{rpc} = \frac{1}{K} \sum_{k=1}^{K} (e^{1-\frac{1}{r_k}} \times L_2(\mathbf{p}_k^1, \mathbf{p}_k^2) - \lambda \times r_k), \tag{10}$$

where $L_2(\cdot)$ denotes L2 distance and the λ is the weight for the regularization term preventing the estimated reliability from zero.

3.4. Training and inference

For the follow-up training and inference, we need to merge the activation maps $\mathbf{A}$ into foreground-background probability map $\widetilde{\mathbf{Y}} \in \mathbb{R}^{2\times H\times W}$. In our implementation, we simply treat the summation of the first half of the activation maps as foreground probability (corresponding the summation of the 1^{st} to the $\frac{K}{2}^{th}$ activation maps), and the summation of the other half as background probability.

Besides, considering the large size and shape variation of the mitochondria, the mito prototype learning may focus on the same (*e.g.*, the entire foreground), making the reliable prototype selection module degeneration. Therefore, we impose a diversity loss to expand the discrepancy among mito features. Formally,

$$\mathcal{L}_{div} = \sum_{i=1}^{K} \sum_{j=1, i\neq j}^{K} \left(\frac{\langle \widetilde{\mathbf{f}}_i, \widetilde{\mathbf{f}}_j \rangle}{\|\widetilde{\mathbf{f}}_i\|_2 \|\widetilde{\mathbf{f}}_j\|_2}\right). \tag{11}$$

The intuition behind this loss is trivial. If the i^{th} and j^{th} features give a high attention to the same region, the $\mathcal{L}_{div}$ will be large and prompt these features to adjust themselves adaptively.

During Training, we calculate the supervise loss for the output prediction of two branch with labeled data by

$$\mathcal{L}_{sup} = \frac{1}{N} \sum_{i=1}^{N} (\mathrm{CE}(\widetilde{\mathbf{Y}}^{\mathbf{1}}, \mathbf{Y}) + \mathrm{CE}(\widetilde{\mathbf{Y}}^{\mathbf{2}}, \mathbf{Y})), \tag{12}$$

where $\mathrm{CE}(\cdot)$ denotes the standard cross entropy loss. As a result, our DualRel is trained by minimizing the overall objective as follows:

$$\mathcal{L} = \mathcal{L}_{sup} + \mathcal{L}_{rpc} + \lambda_{div} \times \mathcal{L}_{div}, \tag{13}$$

where λ_{div} is the trade-off weight.

During inference, we get the predicted mask $\widehat{\mathbf{Y}}$ by apply argmax operation on probability map $\widetilde{\mathbf{Y}}$ from one branch, without extra computation cost.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on the most popular benchmarks including Lucchi [19], Lucchi++ [2], MitoEM dataset [35]. **Lucchi** [19] contains two sub-volumes, acquired from the CA1 hippocampus region of the mouse brain for training and testing. Each sub-volume consists of $165 \times 1024 \times 768$ EM images provided with manual mitochondria segmentation mask. And **Lucchi++** [2] is re-annotated by three neuroscience and biology experts, and has the same size as Lucchi with more accurate mitochondria segmentation mask. **MitoEM dataset** [35] contains two volumes attained from rat and human cortex, dubbed mito-R and mito-H respectively. Each volume consists of 400 training images, 100 validation images and 500 test images with a resolution of 4096×4096. Among them, the ground-truth labels are publicly available for the training set and the validation set. In our experiments, we train on training set and evaluate on validation set.

Evaluation Metrics. We adopt Dice similarity coefficient (DSC) and Jaccard-index coefficient (JAC) to evaluate the accuracy of segmentation in our experiments.

Table 1. Comparison with state-of-the-art re-implemented methods on Lucchi [19] and Lucchi++ [2] dataset under different partition protocols. Partition protocols denote the ratio of labeled data used for training, followed by the actual number of mitochondrial images.

	1/32 (5)		1/16 (10)		1/8 (20)		1/2 (82)	
Method	JAC	DSC	JAC	DSC	JAC	DSC	JAC	DSC
				Lucchi				
MT[NIPS17] [31]	71.85	82.53	72.48	82.83	75.49	84.72	78.60	86.73
CCT[CVPR20] [25]	84.75	91.62	85.48	91.73	85.84	91.94	86.60	92.73
GCT[ECCV20] [11]	83.51	90.93	84.64	92.02	85.86	91.72	86.20	92.41
CPS[CVPR21] [3]	84.55	91.57	84.61	91.18	85.16	91.57	85.51	92.03
DualRel	**85.63**	**92.24**	**86.35**	**92.42**	**87.21**	**93.16**	**87.62**	**93.31**
				Lucchi++				
MT[NIPS17] [31]	79.51	87.11	82.89	90.17	86.91	92.33	87.24	93.42
CCT[CVPR20] [25]	87.36	92.63	87.84	93.22	88.54	93.74	89.36	94.19
GCT[ECCV20] [11]	86.60	92.16	86.95	92.05	87.86	93.12	88.69	93.79
CPS[CVPR21] [3]	86.36	92.42	87.22	93.07	87.87	93.36	88.30	93.52
DualRel	**87.81**	**93.42**	**88.22**	**93.74**	**89.91**	**94.63**	**90.54**	**95.52**

4.2. Implementation Details

In the semi-supervised setting, we sample a certain ratio from the training set as labeled set D^L, and the rest are treat as unlabled set D^U. For a fair comparison, we follow the following guidelines in training and inference stage for all re-implemented methods. We adopt ResNet50 [8] as backbone network for feature extraction by removing the global average pooling (GAP) layer and fully connected layer, and utilize U-Net structure [28] for upsampling. During training, we feed the network with 512×512 images randomly cropped from original EM images with random mirror, random rotate and elastic transform augmentation. Extra color jitter is adopted to reduce overfitting. We use mini-batch SGD to train our model with momentum set to 0.9 and weight decay fixed as 0.0005. We initialize the learning rate with 5×10^{-3} with batch size of 4, and halve the learning rate at the 50%, 70% and 90% of the overall training epoch. During inference, We adopt a sliding window of size 512×512 and step 256.

4.3. Comparison with State-of-the-art Methods

We reproduce the most representative and competitive methods [3, 11, 25, 31] in the semi-supervised natural image semantic segmentation task, and report their performance on Lucchi derivatives (*i.e.*, Lucchi and Lucchi++) and MitoEM dataset, establishing a semi-supervised mitochondria segmentation benchmark.

Lucchi derivatives. Tab. 1 shows the comparison of our method with the state-of-the-art re-implemented methods on Lucchi and Lucchi++ dataset. We consistently observe that our DualRel outperforms all other methods under all partition protocols, which strongly proves the effectiveness of our method. Specifically, our approach achieves 89.9% JAC, 94.6% DSC on Lucchi++ and 87.2% JAC, 93.1% DSC

Table 2. Comparison with state-of-the-art re-implemented methods on MitoEM [35] dataset under different partition protocols.

	1/32 (12)		1/16 (25)		1/8 (50)		1/2 (200)	
Method	JAC	DSC	JAC	DSC	JAC	DSC	JAC	DSC
				mito-R				
MT[NIPS17] [31]	81.11	87.52	84.64	90.61	84.91	90.82	86.99	92.41
CCT[CVPR20] [25]	88.91	93.22	89.48	93.88	89.71	94.03	89.74	94.12
GCT[ECCV20] [11]	88.22	92.81	88.61	92.73	89.03	93.41	89.11	93.72
CPS[CVPR21] [3]	88.71	93.71	89.62	94.41	89.71	94.52	90.02	94.72
DualRel	**89.61**	**94.52**	**90.31**	**94.93**	**90.72**	**95.11**	**90.91**	**95.23**
				mito-H				
MT[NIPS17] [31]	81.90	89.32	82.71	89.62	83.57	90.36	84.63	91.14
CCT[CVPR20] [25]	83.42	90.61	84.18	90.75	84.87	91.24	85.07	91.37
GCT[ECCV20] [11]	82.12	89.81	83.32	90.91	84.62	91.02	84.81	91.06
CPS[CVPR21] [3]	84.22	91.81	84.29	91.47	85.42	92.12	86.31	92.65
DualRel	**85.61**	**92.22**	**85.93**	**92.44**	**86.21**	**92.43**	**86.53**	**92.72**

on Lucchi under $1/8$ partition protocol (only 20 labeled images). This means that even if we only label a small proportion of images, we can train a satisfactory segmentation model in an appropriate way. Compared to the best method CPS [3] in the field of natural image, our DualRel achieves a large margin of 2.05%/1.59% in JAC/DSC under 1/8 partition protocol, which favorably manifests the benefits of prototype-level consistency regularization.

MitoEM dataset. In Tab. 2 we report the performance on MitoEM dataset, which is much more complex than Lucchi and Lucchi++, with larger image size, more messy background and more diverse mitochondrial morphology. DualRel also achieves convincing performance even when the labeled data is scarce. Under 1/32 partition protocol, *i.e.*, training the network with solely 12 labeled images and 388 unlabeled images, we obtain 89.6% JAC, 94.5% DSC on mito-R and 85% JAC, 91.9% DSC on mito-H. This demonstrates the the stability of our method, which can model the reliability of the prototype for a more robust prototype-level supervision, even in the face of more complex scene.

Comparison with fully supervised methods. Tab. 3 tabulates the quantitative results compared with fully supervised methods. We observe that with only 5 annotated images, DualRel even surpasses fully supervised methods that utilize all annotated samples, such as 2D U-Net [28], by 0.7%/1.2% in JAC/DSC. When the number of available labeled samples increases (*e.g.*, half the total images), our method is on par with state-of-the-art fully supervised methods (*e.g.*, Casser [2]) without any post-processing. This sheds light on the great promise of semi-supervised mitochondria segmentation tasks, where light labeling cost can yield competitive performance.

4.4. Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on Lucchi dataset with ResNet50 as backbone under 1/8 partition protocol to analyze each component of our DualRel, including the reliable pixel aggregation module (RPiA), and the reliable prototype selection module (RPrS). Note that we remove all modules except the two branches of feature extractor and upsampling module and a separate cross-attention mechanism, to directly construct prototype-level consistency regularization as our baseline.

Table 3. Comparison with state-of-the-art fully supervised methods on Lucchi dataset [19]. "Post-Processing" stands for whether to use post-processing such as Z-Filtering.

Method	Post-Processing	labels	JAC	DSC
Lucchi [18]	✗	165	75.5	86.0
Peng and Yuan [27]	✗	165	83.3	90.9
2D U-Net [28]	✗	165	84.4	91.5
Cheng (2D) [4]	✗	165	86.5	92.8
Liu [17]	✓	165	86.4	92.6
Khadangi [13]	✗	165	86.5	92.7
Casser [2]	✓	165	88.4	93.8
DualRel (1/32)	✗	5	85.6	92.2
DualRel (1/2)	✗	82	87.6	93.3

Table 4. Evaluation of the effectiveness of different components on Lucchi dataset.

		Pixel-level		
		Threshold	JAC	DSC
CPS [3]		0	85.16	91.57
		0.5	85.20	91.62
		0.7	84.07	91.48
		0.9	83.93	90.89
		Prototype-level		
RPiA		RPrS	JAC	DSC
w/o ref.	w/ ref.			
✓			85.97	91.94
	✓		86.57	92.72
	✓	✓	87.21	93.16

Analysis of the Pixel-level Consistency Regularization. In Sec. 1, we intuitively demonstrate that there exits gap between natural images and mitochondrial images, and directly employing pixel-level consistency regularization on mitochondrial images will inevitably increase the risk of unreliability. Naively, we can select high-confidence pixels based on a pre-defined threshold to attempt to alleviate the unreliability. The upper part of Tab. 4 tabulates the impact of different thresholds on the semi-supervised model based on pixel-level supervision. We observe that selecting high confidence points brings negligible improvement and even impairs model performance when the threshold is set too high. Then, we further quantitatively analyze the reasons for this case from the perspective of expected calibration error (ECE) [7], which measures the discrepancy between the confidence and accuracy of the network output,

Table 5. Comparison with state-of-the-art methods in model calibration and study on different network configurations.

(a) Quantification of expected calibration error (ECE).

Method	ECE (↓) (‰)
MT	15.2
CCT	12.5
GCT	10.2
CPS	8.3
DualRel	1.8

(b) Ablation on different architecture.

	JAC	DSC
CNN	86.99	92.65
Transformer	87.21	93.16

(c) Performance comparison with and without diversity loss.

	JAC	DSC
w/o div.	86.57	92.87
w/ div.	87.21	93.16

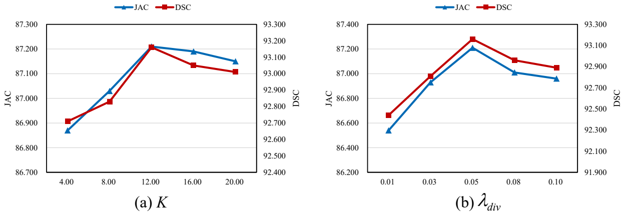
as known as calibration, has been explored. Poorer network calibration, weaker the correlation between the confidence and accuracy of the network output. In other words, even if the network predicts with high confidence, there is a high probability that it will be unreliable. Tab. 5a shows the calibration of different methods, the larger the value, the poorer the calibration. Therefore, it is impracticable to select reliable points only conditioned on confidence.

Analysis of the Prototype-level Consistency Regularization. As shown in Tab. 4, the baseline that utilizes prototype-level supervision achieves a clear lead, compared to methods based on pixel-level consistency regularization. Moreover, Tab. 5a shows our DualRel achieves better calibration benefiting from RPiA and RPrS module, which is in line with the design idea of prototype-level supervision.

Effectiveness of the Reliable Pixel Aggregation Module. As shown in Tab. 4, the introduction of the referential correlation (ref.) in RPiA achieves a certain performance lift compared with the baseline (*i.e.*, a separate cross-attention mechanism without ref.), that is, 91.94% *vs*. 92.72% in DSC. The improvement can be mainly ascribed to the strong ability of the RPiA to rectify the direct pairwise correlation, enabling the prototype absorbs counterpart reliable pixels with the same semantic pattern during the interaction with the pixels.

Effectiveness of the Reliable Prototype Selection Module. The addition of RPrS also contributes to a remarkable performance (86.57% *vs*. 87.21%). This proves the necessity of suppressing the equivocal prototypes, our RPrS can further evaluate the reliability of prototypes in constructing prototype-level consistency regularization.

Effectiveness of Transformer-based Cross-attention Mechanism. We perform the ablation study that replaces the Transformer-like cross-attention by using convolutional networks with similar parameters in Tab. 5b. In detail, We use the convolutional layers followed by a softmax function to change the channel size of the feature map to obtain K mito features when replacing cross attention. The feed-forward network (FFN) is replaced by the convolutional layers, and the rest of the model remains the same. We observe that cross-attention achieves better performance thanks to the long-range modeling capabilities of the transformer.



(a) K (b) λ_{div}

Figure 3. Evaluation of the number of prototypes K, and the hyperparameters λ_{div}.

Effectiveness of Diversity Loss. As shown in Tab. 5c, with the utilization of diversity loss, further improvements can be observed. Diversity loss prevents the prototypes from focusing on similar local semantic clues. And diverse prototypes can capture mitochondria variations and achieve more precise segmentation.

Hyperparameter Evaluations. As shown in Fig. 3, we evaluate how K and λ_{div} affects our model learning. we can observe that the performance continues to grow until $K = 12$, which means it is sufficient to mine 12 semantic patterns of mitochondria. And λ_{div} controls the relative importance of the diversity loss, our model achieves much better performance when $\lambda_{div} = 0.05$.

4.5. Vasualization

Visualization of Predictions. To further analyze and understand the proposed method, we visualize a series of segmentation results and prototype activation areas. As shown in Fig. 4, it can be noticed that other methods tend to incorrectly segment the non-target region (dyed in blue) or are unable to activate all the mitochondria (dyed in red). We deem the main reason is that numerous ambiguous pixels inherent in EM images severely confuse the models designed for natural images. With the assistance of the RPiA, the negative effects of unreliable pixels are mostly eliminated. Besides, the RPrS assigns larger weights to the more reliable prototypes to obtain robust prototype-level supervision. Our DualRel generates more accurate prediction masks compared with other methods, which demonstrates the effectiveness of the cooperation of the above two modules.

Visualization of Correlation Weights. To more intuitively demonstrate the unreliable pixels suppression ability endowed by RPiA, we conduct qualitative visualization on the correlation maps between the mito filters and the image features. As shown in Fig. 6, we find that in the absence of RPiA, pixel features that with distinct semantics sometimes share high similarity with specific prototypes as shown in the parts marked by the red box. We claim that this is detrimental because aggregating features that not align to the

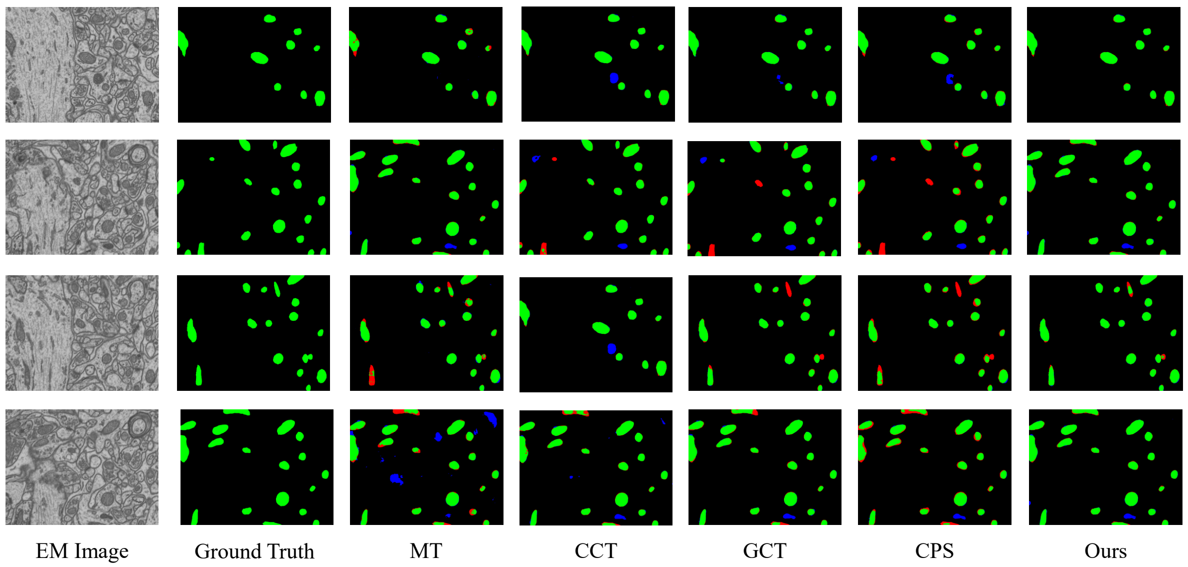


Figure 4. Qualitative results of the proposed DualRel and other methods under 1/8 ratio on Lucchi.

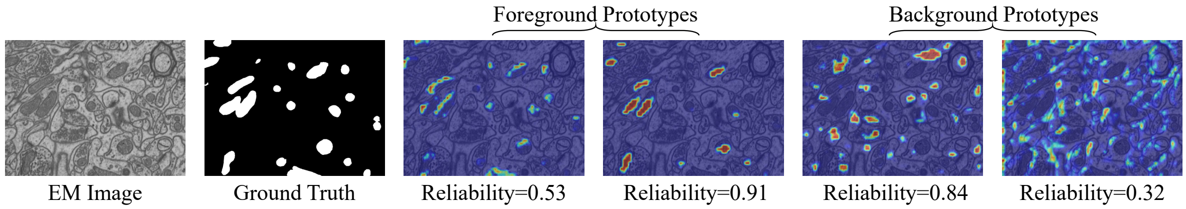


Figure 5. The visualization of the reliability of prototypes and its corresponding activation maps.

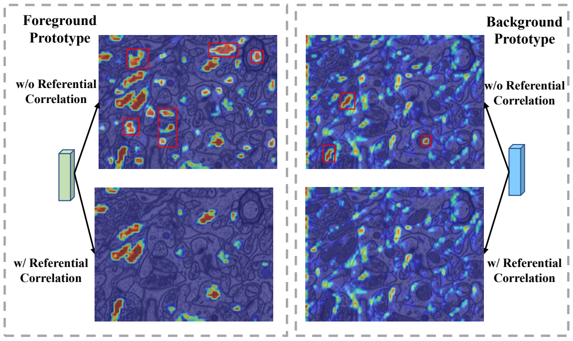

Foreground Prototype
w/o Referential Correlation
w/ Referential Correlation
Background Prototype
w/o Referential Correlation
w/ Referential Correlation

Figure 6. Visualization of the pixel-prototype correlation weights.

prototypes with large weights will lead to confusion. The second row presents the results with RPiA, we observe a significant reduction in false activation, demonstrating the effectiveness of RPiA in suppressing false matches.

Visualization of Reliability and Activation Maps. To vividly present the working mechanism of reliable prototype selection module (RPrS), we visualize the reliability of prototypes and its corresponding activation maps. As shown in Fig. 5, some prototypes with clear foreground or background semantic clues occupy larger weights (*i.e.*, 4^{th} and 5^{th} columns), while some prototypes focusing on mitochondrial boundary patterns are assigned with smaller weights (*i.e.*, 3^{rd} and 6^{th} columns). This is in line with the design idea of RPrS, that is, pursuing implicitly learning the reliability of each prototype in a data-driven way. In this case, the equivocal prototypes will be suppressed while the reliable ones are highlighted in the supervision signals.

5. Conclusion

In this paper, we rethink how to achieve effective semi-supervised mitochondria segmentation. We propose a novel end-to-end dual-reliable (DualRel) network, including a reliable pixel aggregation module and a reliable prototype selection module. Extensive experimental results demonstrate the effectiveness. This work is expected to open a new venue for future research in this field.

6. Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62021001), National Defense Basic Scientific Research Program (Grant JCKY2022911B002).

References

[1] Silvia Campello and Luca Scorrano. Mitochondrial shape changes: orchestrating cell pathophysiology. *EMBO reports*, 11(9):678–684, 2010. 1

[2] Vincent Casser, Kai Kang, Hanspeter Pfister, and Daniel Haehn. Fast mitochondria detection for connectomics. In *Medical Imaging with Deep Learning*, pages 111–120. PMLR, 2020. 1, 5, 6

[3] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2613–2622, 2021. 1, 3, 5, 6

[4] Hsueh-Chien Cheng and Amitabh Varshney. Volume segmentation using convolutional neural networks with limited training data. In *2017 IEEE international conference on image processing (ICIP)*, pages 590–594. IEEE, 2017. 6

[5] MR Duchen. Mitochondria and ca2+ in cell physiology and pathophysiology. *Cell calcium*, 28(5-6):339–348, 2000. 1

[6] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 3

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 6

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 5

[9] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. *arXiv preprint arXiv:1802.07934*, 2018. 3

[10] Wataru Iwasaki, Tsukasa Fukunaga, Ryota Isagozawa, Koichiro Yamada, Yasunobu Maeda, Takashi P Satoh, Tetsuya Sado, Kohji Mabuchi, Hirohiko Takeshima, Masaki Miya, et al. Mitofish and mitoannotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular biology and evolution*, 30(11):2531–2540, 2013. 1

[11] Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pages 429–445. Springer, 2020. 3, 5, 6

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[13] Afshin Khadangi, Thomas Boudier, and Vijay Rajagopal. Em-net: Deep learning for electron microscopy image segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 31–38. IEEE, 2021. 6

[14] Ritwik Kumar, Amelio Vázquez-Reina, and Hanspeter Pfister. Radon-like features and their application to connectomics. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 186–193. IEEE, 2010. 1, 2

[15] Hong Joo Lee, Jung Uk Kim, Sangmin Lee, Hak Gu Kim, and Yong Man Ro. Structure boundary preserving segmentation for medical image with ambiguous boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4817–4826, 2020. 2

[16] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013. 4

[17] Jing Liu, Linlin Li, Yang Yang, Bei Hong, Xi Chen, Qiwei Xie, and Hua Han. Automatic reconstruction of mitochondria and endoplasmic reticulum in electron microscopy volumes by deep learning. *Frontiers in neuroscience*, 14:599, 2020. 6

[18] Aurélien Lucchi, Yunpeng Li, and Pascal Fua. Learning for structured prediction using approximate subgradient descent with working sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1987–1994, 2013. 1, 2, 6

[19] Aurélien Lucchi, Kevin Smith, Radhakrishna Achanta, Graham Knott, and Pascal Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2):474–486, 2011. 2, 5, 6

[20] Lee J Martin. Biology of mitochondria in neurodegenerative diseases. *Progress in molecular biology and translational science*, 107:355–415, 2012. 1

[21] Guanliang Meng, Yiyuan Li, Chentao Yang, and Shanlin Liu. Mitoz: a toolkit for animal mitochondrial genome assembly, annotation and visualization. *Nucleic acids research*, 47(11):e63–e63, 2019. 1

[22] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019. 3

[23] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 3

[24] Philip Newsholme, Celine Gaudel, and Maurico Krause. Mitochondria and diabetes. an intriguing pathogenetic role. *Advances in mitochondrial medicine*, pages 235–247, 2012. 1

[25] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2020. 3, 5, 6

[26] Ismail Oztel, Gozde Yolcu, Ilker Ersoy, Tommi White, and Filiz Bunyak. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1195–1200. IEEE, 2017. 3

[27] Jialin Peng and Zhimin Yuan. Mitochondria segmentation from em images via hierarchical structured contextual forest. *IEEE Journal of Biomedical and Health Informatics*, 24(8):2251–2259, 2019. 1, 2, 6

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmen-

tation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3, 5, 6

[29] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696, 2017. 3

[30] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10938–10947, 2021. 1

[31] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 3, 5, 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[33] Yuan Wang, Rui Sun, Zhe Zhang, and Tianzhu Zhang. Adaptive agent transformer for few-shot segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*, pages 36–52. Springer, 2022. 1

[34] Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4248–4257, 2022. 3

[35] Donglai Wei, Zudi Lin, Daniel Franco-Barranco, Nils Wendt, Xingyu Liu, Wenjie Yin, Xin Huang, Aarush Gupta, Won-Dong Jang, Xueying Wang, et al. Mitoem dataset: large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 66–76. Springer, 2020. 3, 5, 6

[36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3