Alleviate and Mining: Rethinking Unsupervised Domain Adaptation for Mitochondria Segmentation from Pseudo-Label Perspective

Yujia Chen[1,2*], **Rui Sun**[1*], **Wangkai Li**[1], **Huayu Mai**[1], **Naisong Luo**[1], **Yuwen Pan**[1], **Tianzhu Zhang**[1,2†]

[1]Deep Space Exploration Laboratory, University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
{yujia_chen, issunrui, lwklwk, mai556, lns6, panyw}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn

Abstract

Mitochondria segmentation from electron microscopy (EM) images plays a crucial role in biological and medical research. However, models trained on source domains often suffer from performance degradation when applied to target domains due to domain shift. Unsupervised domain adaptation (UDA) methods have been proposed to address this issue, but they often overlook the reliability of pseudo-labels and the effectiveness of supervision signals. In this paper, we propose R4MITO, a novel UDA framework for robust mitochondria segmentation. First, we introduce Reliable Prototype Pseudo-labels to mitigate the inconsistency of class-level features between across domains by leveraging source prototypes to model target prototypes. Second, we devise Correlation-wise Consistency Regularization to exploit inter-pixel correlations, aligning agent-level correlations under various perturbations. Third, we propose Rank-aware Relationship Consistency Regularization to fully utilize the rich information encoded in inter-agent relationships by imposing rank-aware constraints on agent-ranking probability distributions. Extensive experiments on multiple EM datasets demonstrate the superiority of our R4MITO over existing state-of-the-art UDA methods for mitochondria segmentation.

Introduction

Mitochondria are essential cellular organelles maintaining cellular functions (Picard et al. 2011). Automatic mitochondria segmentation (Ascoli 2002; Donohue and Ascoli 2011) is crucial in biological and medical research. Deep learning-based segmentation methods (Çiçek et al. 2016; Oztel et al. 2017; Sun et al. 2021; Wang et al. 2022; Pan et al. 2023; Luo et al. 2023; Sun et al. 2023b; Luo et al. 2024; Li et al. 2025) have advanced significantly, albeit requiring extensive labeled data. Models trained on source domains tend to suffer severe performance degradation when used directly on target domain due to distribution discrepancies from equipment, sample, and parameter variations. While manual target dataset annotation resolves this, it remains impractical due to resource constraints. Bridging this domain gap for improved model generalization presents a significant challenge.

In this work, we focus on the unsupervised domain adaptation task for mitochondria segmentation (UDAMS), which aims to adapt a model trained on source domain equipped with segmentation annotations to target domain in the absence of accessible labels. Existing methods can be broadly categorized as adversarial learning and consistency regularization. Adversarial learning (Huang et al. 2022; Sun et al. 2023a) leverage adversarial learning to learn domain-invariant information. However, this approach often tends to suffer from training instability and convergence challenges. Recently, consistency regularization (Du et al. 2019; Huang et al. 2020; Yin et al. 2023) dominate this feld credited to its simplicity yet competitive performance. The core idea is to align features or outputs of the same image under different perturbations, using a weakly augmented branch as a pseudo-label for the strongly augmented counterpart.

After an in-depth analysis of the consistency regularization paradigm, we argue that pseudo-labels are crucial for UDAMS, aligning with the intuitive understanding derived from the task's definition; pseudo-labels play a dual role - minimizing the inter-domain discrepancy and maximizing the target domain's perturbation consistency. Nonetheless, our analysis reveals that pseudo-labels become a performance bottleneck due to two key inadequately addressed components in previous works: (1) Reliability of pseudo-labels. Without target annotations, models trained solely on the source domain with significant domain shift are prone to suffer from limited coverage of the underlying target class features (*e.g.,* Figure 1 (a) shows source prototypes are not suitable for direct application to the target domain, due to incorrect decision boundary). Furthermore, Lack of labeled target data can also lead to error accumulation, severely hampering the model's ability to generalize effectively. Therefore, it is highly desirable to mitigate the impact of noisy pseudo-labels and ensure that the truly reliable ones are highlighted for enhance efficacy and utility. (2) Effectiveness of supervision signals from pseudo-labels. In order to fully explore the essential information within the images, existing methods (Wu et al. 2021; Yin et al. 2023) construct supervision signals through pixel-level consistency regularization strategy, which aligns intermediate features and outputs, thereby mitigating some domain discrepancies (Figure 1 (c), left). However, independently treating each pixel and heavily relying on strong i.i.d. assumptions potentially

*These authors contributed equally.
†Corresponding author.

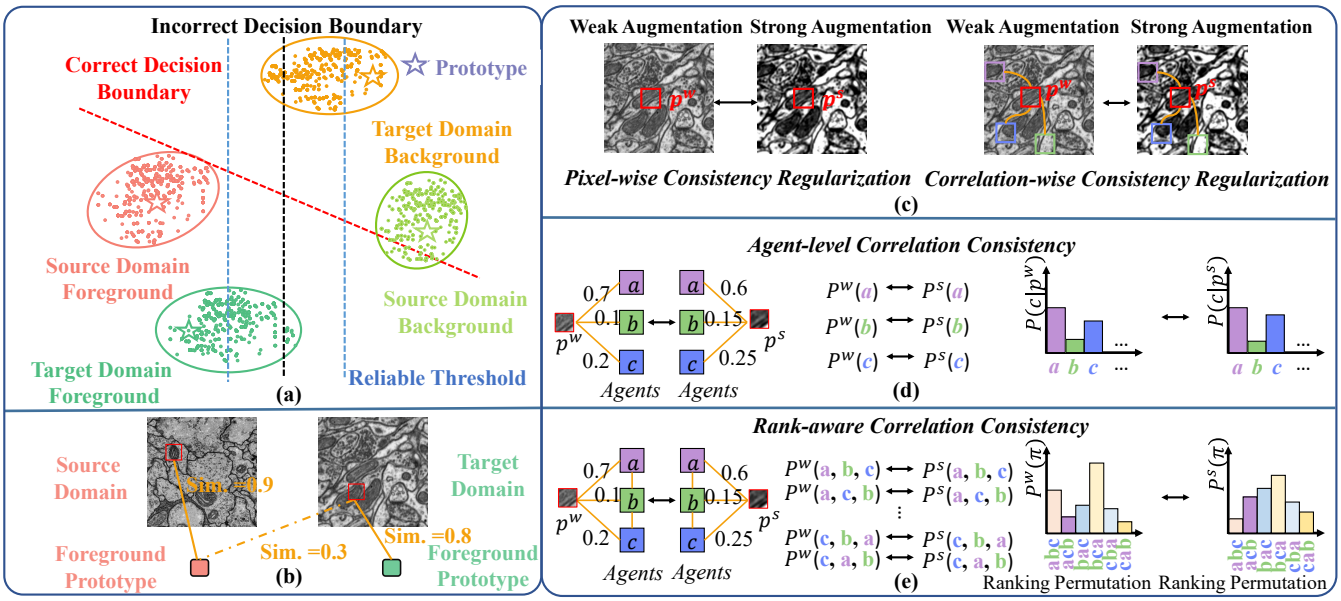

Figure 1: (a) illustrates that utilizing only source domain prototypes can lead to incorrect decision boundaries in the target domain, whereas generating prototypes from high-confidence target features can result in correct decision boundaries. (b) The similarity between pixels and prototypes of the same class is high within the same domain and low across domains. (c) depicts the differences between independent pixel-wise consistency regularization and correlation-wise consistency regularization. (d) shows the straightforward implementation of correlation-wise consistency regularization. (e) illustrates rank-aware correlation consistency. We harness the inter-agent relationship by considering every possible agent rank permutation probability.

misses optimization opportunities.

In this paper, we analyze the limitations of previous methods from the perspective of pseudo-labels, and explore the potential of leveraging inter-pixel correlations to construct more reliable and effective supervision signals for UDAMS. Based on the Gestalt law (Koffka 2013) of intra-class pixel similarity (Fig. 1(b)), we propose **Reliable Prototype Pseudo-labels** to address domain shift-induced feature inconsistency. The method generates target-aware prototypes from high-confidence features of initial source-prototype segmentation, improving pseudo-label reliability.

Intuitively, most existing methods overlook the fact that dense pixel prediction tasks inherently possess rich inter-pixel information beyond basic individual pixel-wise consistency. Inspired by affinity learning (Lee et al. 2017), we note that the remarkable success of affinity learning suggests that the relationships between pixels typically encapsulate far richer information compared to considering individual pixels in isolation. Therefore, we devise **Correlation-wise Consistency Regularization** which is synergistic collaboration between inter-pixel correlation and consistency regularization. Specifically, We model the inter-pixel correlation using a set of representative reference points (referred to as agents) and align the agent-level correlation (*i.e.,* a likelihood vector) of each pixel with the agents under various perturbations, as illustrated in Figure 1 (c) and Figure 1 (d). In essence, the agent-level correlation reflects the consensus among representative agents with a broader receptive field, thus encoding a higher-order consistency regularization. However, it is crucial to obtain suitable agents that encapsulate diverse semantic cues from the original image. Therefore, we select the most representative K agents from the feature map, preserving as much critical information as possible from the original pixels. By leveraging the richer description of the data distribution in the agent-level correlation, we can better exploit the unlabeled data.

Further analysis demonstrates that since the agents inherently encapsulate significant semantic cues, relationships among them also encode rich information. For example, agent a represent the foreground, agent b represent the background, and agent c represent the boundary. Thus, the relationship between a and c should be closer than that between a and b. To fully utilize the information inherent in the inter-agent relationships, instead of considering each agent independently, we design **Rank-aware Relationship Consistency Regularization**, as illustrated in Figure 1 (e). Specifically, we argue that the order of agents is not random but follows a fixed permutation. Thus, we define different agent relationship models based on this ranking. For instance, a pixel's relationship with other agents can be defined in terms of ranking probability. The probability of being ranked first for agent a might be 0.7, while it could be 0.1 for agent b. From this perspective, the ranking permutation reflects the relationship of agents with respect to the pixel. In this way, for a given pixel, we consider every possible rank permutation of the agents (*e.g., abc*, *cba*, *etc*.) and transform agent-level correlations into an agent-ranking probability distribution. By imposing rank-aware constraints, we can obtain stronger and more effective supervision signals. Ultimately, we term our final method as R4MITO.

In this work, our contributions can be summarized as follows: (1) We analyze the limitations of existing methods, including the unreliability in prototype construction and the neglect of consistency regularization for inter-pixel relationships. (2) We propose the R4MITO method, which includes reliable prototype pseudo-labels to extract prototypes suitable for the target domain and rank-aware consistency regularization to fully exploit effective information. (3) Comprehensive experiments on four challenging benchmarks demonstrate the favorable performance of our method against state-of-the-art methods for UDAMS.

Related Work

UDA for Semantic Segmentation

Unsupervised domain adaptation for semantic segmentation (UDASS) comprises two main approaches: adversarial learning and pseudo-labeling. Adversarial learning-based methods (Kim and Byun 2020; Hong et al. 2018; Tsai et al. 2019) employ adversarial learning to assist the model in learning representations that are invariant across domains. Pseudo-labeling techniques (Zou et al. 2018; Li, Yuan, and Vasconcelos 2019; Melas-Kyriazi and Manrai 2021; Hoyer, Dai, and Van Gool 2022; Hoyer et al. 2023; Chen et al. 2019; Du et al. 2019; Feng et al. 2023; Wangkai et al. 2023; Mai et al. 2024a, 2023, 2024b; Sun et al. 2024; Li, Sun, and Zhang 2024) align the feature or output space of different perturbations of the same image. However, directly applying UDA methods designed for natural images to the problem of mitochondrial segmentation can result in a decline in segmentation performance. This is due to the high complexity of EM images, similar morphologies between mitochondria and other structures, scarce discriminative features, and the presence of imaging artifacts

UDA for Mitochondria Segmentation

Existing UDAMS approaches follow UDASS paradigm, developing in two directions. In the field of adversarial learning, Advent (Vu et al. 2019) first proposes the UDAMS task. DA-ISC (Huang et al. 2022) building upon the Advent framework, employs an XOR-based method to mitigate inter-slice differences. SAPAN (Sun et al. 2023a) further eliminates long-range structural differences. In the field of Pseudo-labeling, UALR (Wu et al. 2021) proposes an uncertainty-aware model to rectify noisy labels. DAMT-Net (Peng, Yi, and Yuan 2020)then devises a reconstruction decoder to align the encoder features from different domains. CAFA (Yin et al. 2023) further leverages source domain prototypes and discrepancy minimization methods to enhance the effectiveness of transferring domain-invariant knowledge. Although these methods significantly reduce the impact of the domain gap, they have not effectively established a reliable bridge between the source and target domains. Moreover, they have neglected the relational information between pixels, leading to suboptimal results.

Method

Preliminary

The UDAMS aims at learning an accurate segmentation model in the target domain based on labeled source domain EM volume $P^S = \{\mathbf{x}_i^S, \mathbf{y}_i^S\}_{i=1}^N$ and unlabeled target domain EM volume $P^T = \{\mathbf{x}_j^T\}_{j=1}^M$. N and M denote the number of sequential sections in S and T. Given an EM section $\mathbf{x}_i \in \mathbb{R}^{H\times W}$ (omit the superscript S/T for convenience) from either domain, the encoder first extracts its feature map $\mathbf{F}_i \in \mathbb{R}^{c\times h\times w}$, where h, w and c denote the height, width and channel number of the feature map respectively. The supervised loss on the source domain $\mathcal{L}_{sup}$:

$$\mathcal{L}_{sup} = \sum_{i=1}^{N} CE(f(x_i^S), y_i^S), \tag{1}$$

where CE denotes Cross-Entropy loss. Additionally, Pseudo-label methods are utilized to align features and output spaces across different perturbations. Specifically, two different levels of random perturbations are applied to the input target data x_i^T. Here, $aug(\boldsymbol{x}_i^T)$ represents a weak perturbation, which will serve as the pseudo-label for the strongly perturbed version $\mathcal{A}ug(x_i^T)$. Then, We incorporate two consistency losses on the feature level $\mathcal{L}_{cf}$ and the prediction level $\mathcal{L}_{cp}$, respectively:

$$\mathcal{L}_{cf} = \sum_{i=1}^{M} MSE(F_i^{Aug}, F_i^{aug}), \mathcal{L}_{reg} = \sum_{i=1}^{M} CE(y_i^{Aug}, \hat{y}_i^{aug}), \tag{2}$$

where MSE denotes the standard mean squared error loss, the original total loss can be expressed as:

$$\mathcal{L}_{org} = \mathcal{L}_{sup} + \lambda_{cf}\mathcal{L}_{cf} + \lambda_{reg}\mathcal{L}_{reg}. \tag{3}$$

Reliable Prototype Pseudo-labels

Directly using source domain prototypes can lead to subptimal results due to the inherent large domain discrepancies in EM images. Considering the lack of annotations in the target domain, directly generating target domain prototypes is challenging. Nevertheless, there are numerous domain-invariant features shared between the source and target domains. These can be leveraged by using prototypes generated from the source domain as an intermediate bridge to identify reliable prototypes in the target domain. As shown in Figure 3, we first derive the source prototypes $\mathbf{P}^{\mathcal{S}} = \{\mathbf{p}_k^{\mathcal{S}}\}_{k=1}^K$ from the source features with the corresponding labels. The prototypes can be calculated as the centroid of each class in the feature space:

$$p_k^s = \frac{\sum_{h=1}^{h_s}\sum_{w=1}^{w_s} f_{h,w}^s \mathbb{1}[Y_{h,w}^s = k]}{\sum_{h=1}^{h_s}\sum_{w=1}^{w_s} \mathbb{1}[Y_{h,w}^s = k]}, \tag{4}$$

where $\mathbb{1}$ refers to indicator function, $f_{h,w}^s \in \mathbb{R}$ is the source feature vectors, and h_s, w_s is the height and width of the features. k is the index of class number K. Then, regarding the original prototypes $\{\mathbf{p}_k^{\mathcal{S}}\}_{k=1}^K$ as query, the target feature map $\mathbf{F}$ as keys, and the k^{th} correlation map $\mathbf{C}_k$ is given as:

$$c_k = p_k^s \cdot F^{\mathsf{T}}, \tag{5}$$

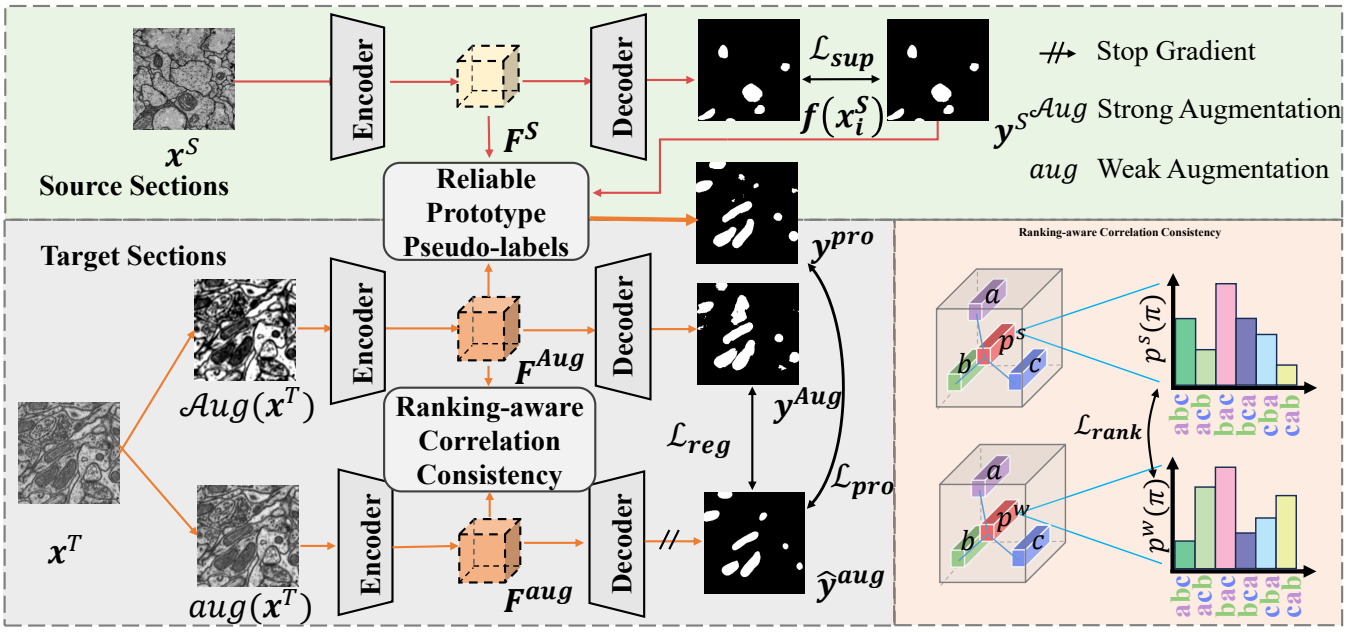

Figure 2: Overview of our method. There are two main modules in R4MITO, i. e., reliable prototype pseudo-labels is responsible for better knowledge transfer between the source and target domains, while ranking-aware correlation consistency aligns more effective semantic information.

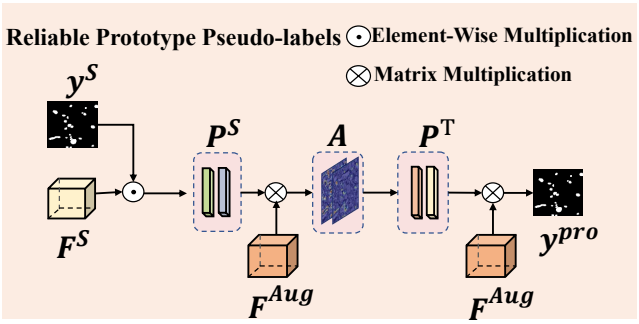

Figure 3: Illustration of Reliable Prototype Pseudo-labels.

where the T refers to the matrix transpose operation. We apply $softmax(\cdot)$ operation along the k dimension and get the initial attention maps $\mathbf{A} = \{\mathbf{a}_k\}_{k=1}^{K}$:

$$[\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K] = softmax([\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_K]), \tag{6}$$

where$[\cdot,\cdot]$ denotes concatenate operation. With the attention maps, We can get masks $\{\mathbf{m}_k\}_{k=1}^{K}$ with high confidence:

$$[\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_K] = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_K] > \tau, \tag{7}$$

where τ is a threshold that needs to be set based on Empirical. $\{\mathbf{m}_k\}_{k=1}^{K}$ constitutes high-confidence pseudo-labels $\widetilde{y^{pro}}$ in the target domain. Subsequently, we can utilize Equation (4) to derive reliable prototypes $\{\mathbf{p}_k^{\mathcal{T}}\}_{k=1}^{K}$ of the target domain. Next, we can achieve prototype-based segmentation results y^{pro} without the need for any additional segmentation head:

$$y_i^{pro} = p_k^{T} \cdot F_i^{Aug\mathsf{T}}, \tag{8}$$

the obtained prototype-based predictions can be supervised using pseudo-labels generated from the outputs of weakly perturbed inputs:

$$\mathcal{L}_{pro} = \sum_{i=1}^{M} CE(y_i^{pro}, \hat{y}_i^{aug}). \tag{9}$$

Correlation-wise Consistency Regularization

Simply enforcing the correlations of all pixels (*i.e.,* the self-correlation matrix) to be consistent is not desirable. Lots of noise in the self-correlation matrix interferes with the optimization of model. Meanwhile, its quadratic complexity will result in a significant computational burden. In order to better model the inter-pixel correlation for consistency regularization, instead of using relationships between all pixels, we select a set of agents. Intuitively, it follows that agents should capture most of the semantic cues. Specifically, we select K features from feature map $\boldsymbol{F} \in \mathbb{R}^{c\times h\times w}$ as agents $\boldsymbol{A} = \{\boldsymbol{f}_i^a\}_{i=1}^{K} \in \mathbb{R}^{K\times c}$. We first obtain the self-attention matrix $s \in \mathbb{R}^{hw\times hw}$ for $\boldsymbol{F}$, calculate s along the pixel dimension, and select the K highest score as agents. This strategy helps in preserving as much critical information as possible in the original pixels. then, we obtain the agent-level correlation $\boldsymbol{c}_{ij} \in \mathbb{R}^{1\times K}$:

$$\boldsymbol{c}_{ij} = softmax(\boldsymbol{f} \cdot \boldsymbol{A}^{\mathsf{T}}), \tag{10}$$

For unlabeled data with different perturbation intensities, we can directly use KL divergence loss to enforce consistency regularization on the agent-level correlation c:

$$\mathcal{L}_{corr} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{HW}\sum_{j=1}^{HW}\ell_{kl}\left(\boldsymbol{c}_{ij}^{Aug}, \boldsymbol{c}_{ij}^{aug}\right). \tag{11}$$

Methods	VNC III → Lucchi (Subset1)				VNC III → Lucchi (Subset2)			
	mAP(%)	F1(%)	MCC(%)	IoU(%)	mAP(%)	F1(%)	MCC(%)	IoU(%)
Oracle	97.5	92.9	92.3	86.8	99.1	94.2	93.7	88.8
NoAdapt	74.1	57,6	58.6	40.5	78.5	61.4	62.0	44.7
DANN (2016)	-	68.2	-	51.9	-	74.9	-	60.1
AdaptSegNet (2018)	-	69.9	-	54.0	-	79.0	-	65.5
Y-Net (2019)	-	68.2	-	52.1	-	71.8	-	56.4
Advent (2019)	78.9	74.8	73.3	59.7	90.5	82.8	81.8	70.7
DAMT-Net (2020)	74.7	60.0	81.3	68.7	94.8	85.8	85.4	75.4
UALR (2021)	80.2	72.5	71.2	57.0	87.2	78.8	77.7	65.2
DA-VSN (2021)	82.8	75.2	73.9	60.3	91.3	83.1	82.2	71.1
DA-ISC (2022)	89.5	81.3	80.5	68.7	92.4	85.2	84.5	74.3
SAPAN (2023a)	91.1	84.1	83.5	72.8	94.4	86.7	86.1	77.1
CAFA (2023)	91.1	83.4	82.8	71.8	94.8	85.8	85.4	75.4
R4MITO(Ours)	**92.6**	**85.4**	**84.3**	**73.7**	**95.0**	**87.9**	**87.5**	**78.6**

Table 1: Comparison with other SOTA methods on the Lucchi dataset. "VNC III → Lucchi(subset1)" means training the model with VNC III as source domain and Lucchi training set as target domain and testing it on Lucchi testing set, and vice versa.

Methods	MitoEM-R → MitoEM-H				MitoEM-H → MitoEM-R			
	mAP(%)	F1(%)	MCC(%)	IoU(%)	mAP(%)	F1(%)	MCC(%)	IoU(%)
Oracle	97.0	91.6	91.2	84.5	98.2	93.2	92.9	87.3
NoAdapt	74.6	56.8	59.2	39.6	88.5	76.5	76.8	61.9
Advent (2019)	89.7	82.0	81.3	69.6	93.5	85.4	84.8	74.6
DAMT-Net (2020)	92.1	84.4	83.7	73.0	94.8	86.0	85.7	75.4
UALR (2021)	90.7	83.8	83.2	72.2	92.6	86.3	85.5	75.9
DA-VSN (2021)	91.6	83.3	82.6	71.4	94.5	86.7	86.3	76.5
DA-ISC (2022)	92.6	85.6	84.9	74.8	96.8	88.5	88.3	79.4
SAPAN (2023a)	93.9	86.1	85.5	75.6	97.0	89.2	88.8	80.6
CAFA (2023)	92.8	86.6	86.0	76.3	96.8	89.2	88.9	80.6
R4MITO(Ours)	**94.9**	**87.5**	**86.7**	**77.3**	**97.2**	**90.5**	**89.8**	**81.8**

Table 2: Comparison with other SOTA methods on the MitoEM dataset. "MitoEM-R → MitoEM-H" means training the model with MitoEM-R training set as the source domain and MitoEM-H training set as the target domain and testing it on MitoEM-H validation set, and vice versa.

Rank-Aware Consistency Regularization

Furthermore, we observe that there exists also a relationship between each agent. Consequently, we design a rank-aware consistency regularization to further enhance the model's learning capability. Specifically, we treat class ranking as a random event rather than a deterministic permutation. In other words, every permutation of the classes exists with some probability, rather than only the permutation from largest to smallest being considered. The probability of one permutation $\pi \in \mathcal{P}$ ($|\mathcal{P}| = K!$) given c can be derived as:

$$P(\pi|\boldsymbol{c}) = \prod_{k=1}^{K} \frac{\boldsymbol{c}_{\pi(k)}}{\sum_{k'=k}^{K} \boldsymbol{c}_{\pi(k')}}, \tag{12}$$

where $\pi(k)$ denotes the k^{th} agent index of this permutation. For example, suppose we have three agents: a, b and c. One permutation of these three agents is $\pi = (a, b, c)$. Based on the agent-level correlation c, we can derive the probability of permutation π :

$$P(\pi|\boldsymbol{c}) = \frac{\boldsymbol{c}(a)}{\boldsymbol{c}(a) + \boldsymbol{c}(b) + \boldsymbol{c}(c)} \cdot \frac{\boldsymbol{c}(b)}{\boldsymbol{c}(b) + \boldsymbol{c}(c)} \cdot \frac{\boldsymbol{c}(c)}{\boldsymbol{c}(c)}. \tag{13}$$

By calculating the probability of each possible order, we model the agent-level correlation c as an agent-ranking probability distribution $P(\mathcal{P}|\boldsymbol{c}) \in \mathbb{R}^{1\times|\mathcal{P}|}$, However, computing the relationships among all agents results in a complexity of $K!$, which is impractical in real scenarios. For computational efficiency, we focus on the permutations of the top-4 agents for each pixel, based on our observation that in every agent-level correlation, the top-4 agents have captured almost all semantic information. The rank-aware correlation consistency regularization can be obtained by:

$$\mathcal{L}_{rank} = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{HW}\sum_{j=1}^{HW}\ell_{kl}\left(P(\mathcal{P}|\boldsymbol{c}_{ij}^{Aug}), P(\mathcal{P}|\boldsymbol{c}_{ij}^{aug})\right). \tag{14}$$

Training and Inference

The total training objective $\mathcal{L}_{total}$ is formulated as :

$$\mathcal{L}_{total} = \mathcal{L}_{org} + \lambda_{pro}\mathcal{L}_{pro} + \lambda_{rank}\mathcal{L}_{rank}, \tag{15}$$

where λ_{pro} and λ_{rank} and are hyperparameters for balancing different terms. During validation, we only adopt the trained feature extractor and decoder to predict the target images.

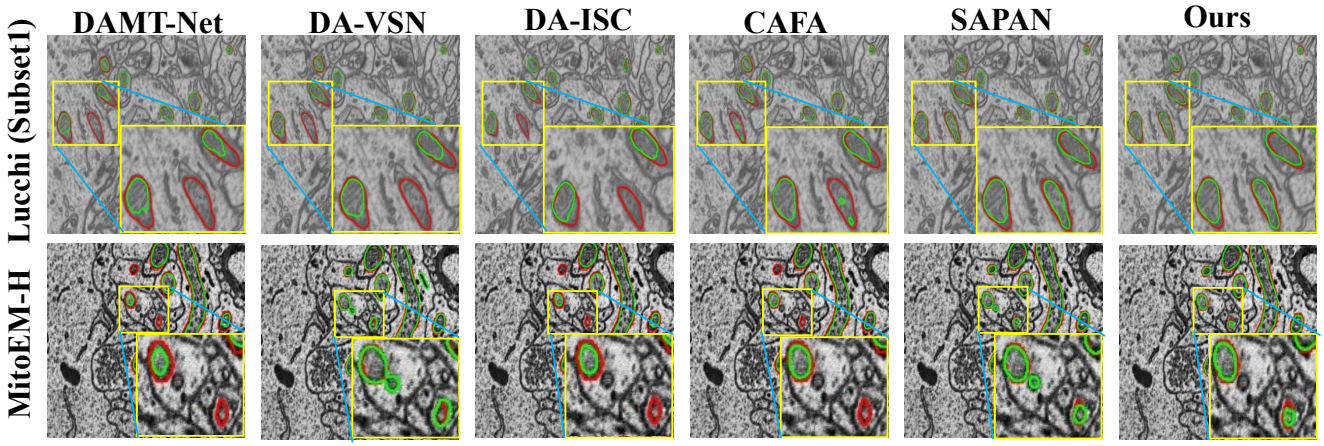

Figure 4: Qualitative comparison with different methods. Note that the red and green contours denote the ground-truth and prediction. And we mark significant improvements using yellow boxes.

Experiments

Datasets

We evaluate our approach on three widely used EM datasets: the VNC III (Gerhard et al. 2013) dataset, the Lucchi dataset (Lucchi, Li, and Fua 2013), and the MitoEM dataset (Wei et al. 2020). These datasets exhibit significant diversity and domain discrepancies, making domain adaptation both challenging and realistic. VNC III Dataset Contains 20 sections (1024 × 1024) from the Drosophila, imaged with serial-section transmission electron microscopy (ssTEM) at a resolution of 50 × 5 × 5 nm³. Lucchi Dataset Consists of two subsets (165 × 1024 × 768) from mouse, imaged with focused ion beam scanning electron microscopy (FIB-SEM) at a resolution of 5 × 5 × 5 nm³. MitoEM Dataset Comprises two subsets (1000 × 4096 × 4096) from rat and human, imaged with multi-beam scanning electron microscopy (mbSEM) at a resolution of 30 × 8 × 8 nm³.

Evaluation Metrics

Following the works (Yin et al. 2023; Wei et al. 2020), four widely used metrics are employed for evaluation, namely, mean Average Precision (mAP), F1 score, Matthews Correlation Coefficient (MCC), and Intersection over Union (IoU).

Implementation Details

The experiments are implemented on PyTorch (Paszke et al. 2017), utilizing an NVIDIA 3090 GPU with 24GB memory. A 4-stage U-Net is selected as the network architecture, with dimensional changes of [64, 128, 256, 512]. Data augmentation methods follow the principles of DA-ISC (Huang et al. 2022), such as flipping, normalization. The augmented EM images are randomly cropped into patches of size 512×512. All models are trained using the Adam optimizer with a batch size of 2, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is initially set at 1×10^{-4}. The balancing weights λ_{pro}, λ_{rank} are set as 0.1 and 0.1, respectively. The reliable target domain feature threshold τ is 0.7, and the number K of agents is 128. The values of λ_{cf} and λ_{reg} in $\mathcal{L}_{org}$ are based on the settings used in CAFA (Yin et al. 2023). The models are trained for a total of 100k iterations.

Reliable Prototype	Correlation	Rank	IoU(%)
			71.8
✓			72.9
	✓		72.2
	✓	✓	72.8
✓	✓	✓	**73.7**

Table 3: Ablation studies of different components.

Comparison with State-of-the-art Methods

In addition to comparing our approach with current state-of-the-art UDA for mitochondrial segmentation, we include two baselines: "Oracle" which refers to directly training the model on the target domain dataset with the ground truth labels, and "NoAdapt" which refers to training the model only using the source domain dataset without the target domain data. Table 1 and Table 2 present the performance comparison between our method and other competitive methods.

We observe that our proposed method outperforms the state-of-the-art method CAFA (Yin et al. 2023) across all metrics and settings. On the VNC III → Lucchi (Subset1) and VNC III → Lucchi (Subset2) setups, the IoU reach 73.7% and 78.6%, respectively. Compared to the current SOTA method CAFA (Yin et al. 2023) the improvement is 1.9% and 3.2%. In the MitoEM-R → MitoEM-H and MitoEM-H → MitoEM-R settings, the IoU reach 77.3% and 81.8%. Compared to the current SOTA method CAFA (Yin et al. 2023) the improvement is 1.0% and 1.2%. It is worth noting that the 'Oracle' method exhibits only a marginal improvement of 1.0% in terms of mAP when transitioning from the MitoEM-H to the MitoEM-R dataset. This indicates that our method possesses a level of supervision capability close to that of the ground truth.

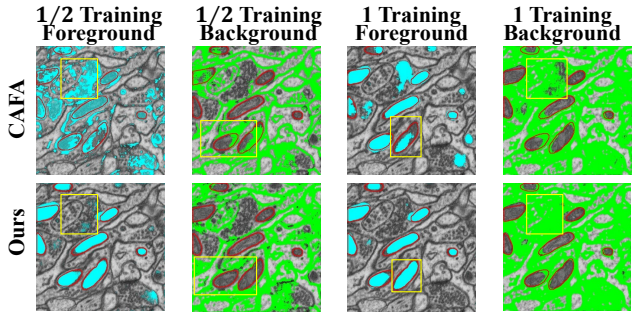

Figure 5: During training, features with similarity greater than τ to different prototypes are visualized.

Strategy	IoU(%)
All	73.4
Random	73.1
TopK	**73.7**

Table 5: Ablation studies of different components.

Agents Num	IoU(%)
32	73.3
64	73.5
128	**73.7**
256	73.6

Table 6: Evaluation of Agents number K.

Ablation Study and Analysis

We conduct comprehensive ablation studies on our proposed method in the VNC III $\rightarrow$ Lucchi(Subset1) setting. We use CAFA (Yin et al. 2023) as our baseline. Table 3 summarizes the results of module ablation studies under different configurations. (1) Compared to the baseline, the reliable prototype pseudo-labels has brought a 1.1% performance improvement. This enhancement can be attributed to it being more suitable for classifying the features of target domain, resulting in more accurate decision boundaries. (2) With the utilization of correlation-wise consistency regularization, IoU is improved by 0.4%, indicating that the relationships between pixel and the agents contain valuable information. (3) By constructing an agent-ranking probability distribution, more effective supervision signals can be generated, which is reflected in further performance improvements of 1.0%. (4) Our method improves upon the baseline by 1.9%, demonstrating its effectiveness. (5) To explore component effectiveness, including top K agents selection strategy, coefficients λ_{pro} and λ_{rank}, agent number and threshold settings, as shown in Tabel 4.

Coefficient				Threshold	
λ_{pro}	IoU(%)	λ_{rank}	IoU(%)	τ	IoU(%)
0.05	73.3	0.05	73.4	0.5	73.2
0.1	**73.7**	0.1	**73.7**	0.6	73.6
0.3	73.6	0.3	73.5	0.7	**73.7**
0.5	73.4	0.5	73.2	0.8	73.6

Table 4: Coefficient λ_{pro} and λ_{rank} and Threshold τ.

Visualization

Visualization of Predictions. To visually assess the effectiveness of our method, we conduct qualitative analysis, as shown in Figure 4. We visualize a series of segmentation results, where red represents the ground truth contour, and green represents the model's predicted contour. It is clear to observe (within the yellow frame) that, compared to other methods, our method has more correctly activated areas and fewer incorrectly activated areas, indicating better segmentation performance.

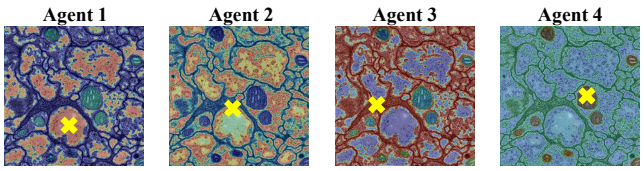

Figure 6: Visualization of agent activation maps. The yellow cross denotes the position of agents in the original image.

Visualization of Reliable Prototype. To demonstrate the effectiveness of reliable prototype, we visualize the pixels predicted to be above the threshold using prototypes, with green representing the foreground, blue representing the background, and red representing the ground truth, As shown in Figure 5. At both the halfway point and the end of the training process, compared to the CAFA (Yin et al. 2023) method, our method exhibits more accurate activations. This indicates that our reliable target domain prototypes more closely match the true decision boundaries of target domain.

Visualization of Agents. In Figure 6, we visualize the significant information represented by different agents. It can be observed that these agents activate different parts of image and resonate favorably with diverse semantic cues derived from the original pixels. From left to right, the visualizations depict the foreground of other organelles, the boundaries of these organelles, the gaps between organelles, and the foreground of mitochondria. These carefully selected agents retain as much critical information as possible from the original image, which facilitates the subsequent construction of correlation consistency.

Conclusion

In this paper, we propose R4MITO, a novel unsupervised domain adaptation framework for robust EM image mitochondria segmentation. Our contributions include: (1) Reliable Prototype Pseudo-labels to mitigate domain shift, establishing a reliable bridge for knowledge transfer between the source and target domains; (2) Correlation-wise Consistency Regularization that exploits inter-pixel correlations through agent-level correlation alignment, exploring more unlabeled supervision signals, and (3) Rank-aware Relationship Consistency Regularization that utilizes inter-agent relationships, further exploring higher-order semantic supervision signals. Extensive experiments demonstrate R4MITO's superiority over state-of-the-art methods for UDAMS, effectively addressing pseudo-label reliability and supervision signal efficacy limitations.

Acknowledgments

This work was supported by National Nature Science Foundation of China (Grant 62021001, 62394354), Youth Innovation Promotion Association.

References

Ascoli, G. A. 2002. *Computational neuroanatomy: Principles and methods*. Springer Science & Business Media.

Chen, C.; Dou, Q.; Chen, H.; Qin, J.; and Heng, P.-A. 2019. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 865–872.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, 424–432. Springer.

Donohue, D. E.; and Ascoli, G. A. 2011. Automated reconstruction of neuronal morphology: an overview. *Brain research reviews*, 67(1-2): 94–102.

Du, L.; Tan, J.; Yang, H.; Feng, J.; Xue, X.; Zheng, Q.; Ye, X.; and Zhang, X. 2019. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 982–991.

Feng, W.; Ju, L.; Wang, L.; Song, K.; Zhao, X.; and Ge, Z. 2023. Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 623–631.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Gerhard, S.; Funke, J.; Martel, J.; Cardona, A.; and Fetter, R. 2013. Segmented anisotropic sstem dataset of neural tissue. *figshare*, 0–0.

Guan, D.; Huang, J.; Xiao, A.; and Lu, S. 2021. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8053–8064.

Hong, W.; Wang, Z.; Yang, M.; and Yuan, J. 2018. Conditional generative adversarial network for structured domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1335–1344.

Hoyer, L.; Dai, D.; and Van Gool, L. 2022. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9924–9935.

Hoyer, L.; Dai, D.; Wang, H.; and Van Gool, L. 2023. MIC: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11721–11732.

Huang, J.; Lu, S.; Guan, D.; and Zhang, X. 2020. Contextual-relation consistent domain adaptation for semantic segmentation. In *European conference on computer vision*, 705–722. Springer.

Huang, W.; Liu, X.; Cheng, Z.; Zhang, Y.; and Xiong, Z. 2022. Domain adaptive mitochondria segmentation via enforcing inter-section consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 89–98. Springer.

Kim, M.; and Byun, H. 2020. Learning texture invariant representation for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12975–12984.

Koffka, K. 2013. *Principles of Gestalt psychology*. routledge.

Lee, K.; Zung, J.; Li, P.; Jain, V.; and Seung, H. S. 2017. Superhuman accuracy on the SNEMI3D connectomics challenge. *arXiv preprint arXiv:1706.00120*.

Li, W.; Sun, R.; and Zhang, T. 2024. A Universal Degradation-based Bridging Technique for Domain Adaptive Semantic Segmentation. arXiv:2412.10339.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6936–6945.

Li, Z.; Wang, Y.; Li, W.; Sun, R.; and Zhang, T. 2025. Localization and expansion: A decoupled framework for point cloud few-shot semantic segmentation. In *European Conference on Computer Vision*, 18–34. Springer.

Lucchi, A.; Li, Y.; and Fua, P. 2013. Learning for structured prediction using approximate subgradient descent with working sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1987–1994.

Luo, N.; Pan, Y.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Camouflaged instance segmentation via explicit decamouflaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17918–17927.

Luo, N.; Sun, R.; Pan, Y.; Zhang, T.; and Wu, F. 2024. Electron microscopy images as set of fragments for mitochondrial segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3981–3989.

Mai, H.; Sun, R.; Wang, Y.; Zhang, T.; and Wu, F. 2024a. Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4162–4170.

Mai, H.; Sun, R.; Zhang, T.; and Wu, F. 2024b. RankMatch: Exploring the Better Consistency Regularization for Semi-supervised Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3391–3401.

Mai, H.; Sun, R.; Zhang, T.; Xiong, Z.; and Wu, F. 2023. Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19617–19626.

Melas-Kyriazi, L.; and Manrai, A. K. 2021. Pixmatch: Unsupervised domain adaptation via pixelwise consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12435–12445.

Oztel, I.; Yolcu, G.; Ersoy, I.; White, T.; and Bunyak, F. 2017. Mitochondria segmentation in electron microscopy volumes using deep convolutional neural network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1195–1200. IEEE.

Pan, Y.; Luo, N.; Sun, R.; Meng, M.; Zhang, T.; Xiong, Z.; and Zhang, Y. 2023. Adaptive template transformer for mitochondria segmentation in electron microscopy images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21474–21484.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Peng, J.; Yi, J.; and Yuan, Z. 2020. Unsupervised mitochondria segmentation in EM images via domain adaptive multitask learning. *IEEE Journal of Selected Topics in Signal Processing*, 14(6): 1199–1209.

Picard, M.; Taivassalo, T.; Gouspillou, G.; and Hepple, R. T. 2011. Mitochondria: isolation, structure and function. *The Journal of physiology*, 589(18): 4413–4421.

Roels, J.; Hennies, J.; Saeys, Y.; Philips, W.; and Kreshuk, A. 2019. Domain adaptive segmentation in volume electron microscopy imaging. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 1519–1522. IEEE.

Sun, R.; Li, Y.; Zhang, T.; Mao, Z.; Wu, F.; and Zhang, Y. 2021. Lesion-aware transformers for diabetic retinopathy grading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10938–10947.

Sun, R.; Mai, H.; Luo, N.; Zhang, T.; Xiong, Z.; and Wu, F. 2023a. Structure-decoupled adaptive part alignment network for domain adaptive mitochondria segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 523–533. Springer.

Sun, R.; Mai, H.; Zhang, T.; and Wu, F. 2024. DAW: exploring the better weighting function for semi-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36.

Sun, R.; Wang, Y.; Mai, H.; Zhang, T.; and Wu, F. 2023b. Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1218–1228.

Tsai, Y.-H.; Hung, W.-C.; Schulter, S.; Sohn, K.; Yang, M.-H.; and Chandraker, M. 2018. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7472–7481.

Tsai, Y.-H.; Sohn, K.; Schulter, S.; and Chandraker, M. 2019. Domain adaptation for structured output via discriminative patch representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1456–1465.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2517–2526.

Wang, Y.; Sun, R.; Zhang, Z.; and Zhang, T. 2022. Adaptive agent transformer for few-shot segmentation. In *European Conference on Computer Vision*, 36–52. Springer.

Wangkai, L.; Zhaoyang, L.; Rui, S.; Huayu, M.; Naisong, L.; Wang, Y.; Yuwen, P.; Guoxin, X.; Huakai, L.; Zhiwei, X.; et al. 2023. Maunet: Modality-aware anti-ambiguity u-net for multi-modality cell segmentation. In *Competitions in Neural Information Processing Systems*, 1–12. PMLR.

Wei, D.; Lin, Z.; Franco-Barranco, D.; Wendt, N.; Liu, X.; Yin, W.; Huang, X.; Gupta, A.; Jang, W.-D.; Wang, X.; et al. 2020. Mitoem dataset: Large-scale 3d mitochondria instance segmentation from em images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 66–76. Springer.

Wu, S.; Chen, C.; Xiong, Z.; Chen, X.; and Sun, X. 2021. Uncertainty-aware label rectification for domain adaptive mitochondria segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 191–200. Springer.

Yin, D.; Huang, W.; Xiong, Z.; and Chen, X. 2023. Class-Aware Feature Alignment for Domain Adaptative Mitochondria Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 238–248. Springer.

Zou, Y.; Yu, Z.; Kumar, B.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, 289–305.