Structure-Decoupled Adaptive Part Alignment Network for Domain Adaptive Mitochondria Segmentation

Rui Sun[1], Huayu Mai[1], Naisong Luo[1], Tianzhu Zhang[1,2,3(✉)], Zhiwei Xiong[1,2], and Feng Wu[1,2]

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

Abstract. Existing methods for unsupervised domain adaptive mitochondria segmentation perform feature alignment via adversarial learning, and achieve promising performance. However, these methods neglect the differences in structure of long-range sections. Besides, they fail to utilize the context information to merge the appropriate pixels to construct a part-level discriminator. To mitigate these limitations, we propose a Structure-decoupled Adaptive Part Alignment Network (SAPAN) including a structure decoupler and a part miner for robust mitochondria segmentation. The proposed SAPAN model enjoys several merits. First, the structure decoupler is responsible for modeling long-range section variation in structure, and decouple it from features in pursuit of domain invariance. Second, the part miner aims at absorbing the suitable pixels to aggregate diverse parts in an adaptive manner to construct part-level discriminator. Extensive experimental results on four challenging benchmarks demonstrate that our method performs favorably against state-of-the-art UDA methods.

Keywords: Mitochondria segmentation · Unsupervised domain adaptation · Electron microscopy images

1 Introduction

Automatic mitochondria segmentation from electron microscopy (EM) volume is a fundamental task, which has been widely applied to basic scientific research and

R. Sun and H. Mai—Equal Contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_50.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 523–533, 2023.
https://doi.org/10.1007/978-3-031-43901-8_50

clinical diagnosis [1,12,13,15,17]. Recent works like [6,7,16,21,22] have achieved conspicuous achievements attributed to the development of deep neural networks [10,23,26,27]. However, these approaches tend to suffer from severe performance degradation when evaluated on target volume (*i.e.*, target domain) that are sampled from a different distribution (caused by different devices used to image different organisms and tissues) compared to that of training volume (*i.e.*, source volume/domain). Therefore, how to alleviate this gap to empower the learned model generalization capability is very challenging.

In this work, we tackle the unsupervised domain adaptive (UDA) problem, where there are no accessible labels in the target volume. To alleviate this issue, representative and competitive methods [5,8,18,19,29,30] attempt to align feature distribution from source and target domains via adversarial training in a pixel-wise manner, that is, learning domain-invariant features against the substantial variances in data distribution. However, after an in-depth analysis of adversarial training, we find two key ingredients lacking in previous works. (1) **Structure-entangled feature.** In fact, there exists a large variation in the complex mitochondrial structure from different domains, as proven in [28]. However, previous methods only focus on domain gap either in each individual section or in adjacent sections, and neglect the differences in morphology and distribution (*i.e.*, structure) of long-range sections, leading to sub-optimal results in the form of hard-to-align features. (2) **Noisy discrimination.** Intuitively, humans can quickly distinguish domain of images with the same categories from cluttered backgrounds by automatically decomposing the foreground into multiple local parts, and then discriminate them in a fine-grained manner. Inspired by this, we believe that during adversarial learning, relying solely on context-limited pixel-level features to discriminate domains will inevitably introduce considerable noise, considering that the segmentation differences between the source and target domain are usually in local parts (*e.g.*, boundary). Thus, it is highly desirable to make full use of the context information to merge the appropriate neighboring pixel features to construct a part-level discriminator.

In this paper, we propose a **S**tructure-decoupled **A**daptive **P**art **A**lignment **N**etwork (**SAPAN**) including a structure decoupler and a part miner for robust mitochondria segmentation. (1) In the **structure decoupler**, we draw inspiration from [25] to model long-range section variation in distribution and morphology (*i.e.*, structural information), and decouple it from features in pursuit of domain invariance. In specific, we first prepend a spatial smoothing mechanism for each pixel in the current section to seek the corresponding location of other sections to attain the smoothed features, which are subsequently modulated to obtain decoupled features with easy-to-align properties. (2) Then, we devise a **part miner** as discriminator, which can dynamically absorb the suitable pixels to aggregate diverse parts against noise in an adaptive manner, thus the detailed local differences between the source and target domain can be accurately discriminated. Extensive experimental results on four challenging benchmarks demonstrate that our method performs favorably against SOTA UDA methods.

2 Related Work

Domain adaptation in EM volume [5,8,18,19,29,30] has attracted the attention of more and more researchers due to the difficulty in accessing manual annotation. In [29], they employ self-training paradigm while in [18] adversarial training is adopted and performs better. To further improve the domain generalization, [5] considering the inter-section gap. However, those methods neglect the differences in morphology and distribution (*i.e.*, structure) of long-range sections. In addition, existing adversarial training [4,5,18] adopt context-limited pixel-wise discriminator leading to sub-optimal results.

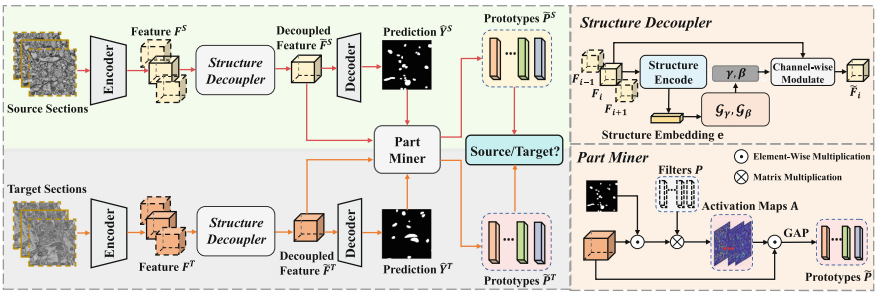

Fig. 1. Overview of our method. There are two main modules in SAPAN, *i.e.*, the structure decoupler for decoupling the feature from the domain-specific structure information and the part miner for adaptively discovering different parts.

3 Method

3.1 Overview

The domain adaptive mitochondria segmentation task aims at learning an accurate segmentation model in the target domain based on a labeled source domain EM volume $V^S = \{\mathbf{X}_i^S, \mathbf{Y}_i^S\}_{i=1}^N$ and an unlabeled target domain EM volume $V^T = \{\mathbf{X}_j^T\}_{j=1}^M$. As shown in Fig. 1, 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 of U-Net [20] first extracts its feature map $\mathbf{F}_i \in \mathbb{R}^{h \times w \times C}$, where h, w and C denote the height, width and channel number of the feature map respectively. A structure decoupler is applied to decouple the extracted feature map from domain-specific structure knowledge, which involves a channel-wise modulation mechanism. Subsequently, the decoupled feature map $\widetilde{\mathbf{F}}_i$ is fed into the decoder of U-Net which outputs prediction of the original size. In the end, we design a part miner to dynamically divide foreground and background into diverse parts in an adaptive manner, which are utilized to facilitate adversarial training. The details are as follows.

3.2 Structure Decoupler

In order to bridge the gap in mitochondrial structure between domains, a structure decoupler is designed to decouple the extracted feature map from domain-specific structure information, realized by a channel-wise modulation mechanism. To better model the structure information, we first apply attention-based spatial smoothing for adjacent sections. Concretely, given feature maps of adjacent sections $\mathbf{F}_i$ and $\mathbf{F}_{i+1}$, we define the spatial smoothing $\mathcal{S}_{i+1}(\cdot)$ w.r.t $\mathbf{F}_{i+1}$ as: each pixel $\mathbf{f}_{i,j} \in \mathbb{R}^{1\times C}$ $(j = 1, 2, ..., hw)$ in feature map $\mathbf{F}_i$ as query, the feature map of adjacent section $\mathbf{F}_{i+1} \in \mathbb{R}^{h\times w\times C}$ as keys and values. Formally,

$$\mathcal{S}_{i+1}(\mathbf{f}_{i,j}) = \text{softmax}(\frac{\mathbf{f}_{\mathrm{i,j}}\mathbf{F}_{\mathrm{i+1}}^{\mathsf{T}}}{\sqrt{\mathrm{C}}})\mathbf{F}_{\mathrm{i+1}}, \tag{1}$$

where the T refers to the matrix transpose operation and the $\sqrt{C}$ is a scaling factor to stabilize training. We compute the structure difference $\mathbf{D}_i \in \mathbb{R}^{h\times w\times C}$ between $\mathbf{F}_i$ and its adjacent $\mathbf{F}_{i+1}$ and $\mathbf{F}_{i-1}$ by:

$$\mathbf{D}_i = [\mathbf{F}_i - \mathcal{S}_i(\mathbf{F}_{i+1})] + [\mathbf{F}_i - \mathcal{S}_i(\mathbf{F}_{i-1})]. \tag{2}$$

The final structure embedding $\mathbf{e} \in \mathbb{R}^{1\times C}$ for each domain is calculated by exponential momentum averaging batch by batch:

$$\mathbf{e}_b = \frac{1}{B}\sum_{i=1}^{B} \text{GAP}(\mathbf{D}_i),\ \ \mathbf{e} = \theta\mathbf{e} + (1-\theta)\mathbf{e}_b, \tag{3}$$

where $\text{GAP}(\cdot)$ denotes the global average pooling, B denotes the batch size and $\theta \in [0, 1]$ is a momentum coefficient.

In this way, $\mathbf{e}^S$ and $\mathbf{e}^T$ condense the structure information for the whole volume of the corresponding domain. To effectively mitigate the discrepancy between different domains, we employ channel-wise modulation to decouple the feature from the domain-specific structure information. Taking source domain data $\mathbf{F}_i^S$ as example, we first produce the channel-wise modulation factor $\gamma^S \in \mathbb{R}^{1\times C}$ and $\beta^S \in \mathbb{R}^{1\times C}$ conditioned on $\mathbf{e}^S$:

$$\gamma^S = \mathcal{G}_\gamma(\mathbf{e}^S), \beta^S = \mathcal{G}_\beta(\mathbf{e}^S), \tag{4}$$

where $\mathcal{G}_\gamma(\cdot)$ and $\mathcal{G}_\beta(\cdot)$ shared by the source domain and target domain, are implemented by two linear layers and an activation layer. Then the decoupled source feature can be obtained by:

$$\widetilde{\mathbf{F}}_i^S = \gamma^S \odot \mathbf{F}_i^S + \beta^S, \tag{5}$$

where $\odot$ denotes element-wise multiplication. The decoupled target feature $\widetilde{\mathbf{F}}_i^T$ can be acquired in a similar way by Eq. (4) and Eq. (5). Subsequently, the structure decoupled feature is fed into the decoder of U-Net, which outputs the foreground-background probability map $\widetilde{\mathbf{Y}}_i \in \mathbb{R}^{2\times H\times W}$. The finally prediction $\widehat{\mathbf{Y}}_i \in \mathbb{R}^{H\times W}$ can be obtained through simply apply $\text{argmax}(\cdot)$ operation on $\widetilde{\mathbf{Y}}_i$.

3.3 Part Miner

As far as we know, a good generator is inseparable from a powerful discriminator. To inspire the discriminator to focus on discriminative regions, we design a part miner to dynamically divide foreground and background into diverse parts in an adaptive manner, which are classified by the discriminator $\mathcal{D}_{part}$ subsequently.

To mine different parts, we first learn a set of part filters $\mathbf{P} = \{\mathbf{p}_k\}_{k=1}^{2K}$, each filter $\mathbf{p}_k$ is represented as a C-dimension vector to interact with the feature map $\widetilde{\mathbf{F}}$ (omit the subscript i for convenience). The first half $\{\mathbf{p}_k\}_{k=1}^{K}$ are responsible for dividing the foreground pixels into K groups and vice versa. Take the foreground filters for example. Before the interaction between $\mathbf{p}_k$ and $\widetilde{\mathbf{F}}$, we first filter out the pixels belonging to the background using downsampled prediction $\widehat{\mathbf{Y}}'$. Then we get K activation map $\mathbf{A}_i \in \mathbb{R}^{h \times w}$ by multiplying $\mathbf{p}_k$ with the masked feature map:

$$\mathbf{A}_i = \text{sigmoid}(\mathbf{p}_\text{k}(\widehat{\mathbf{Y}}' \odot \widetilde{\mathbf{F}})^\mathsf{T}). \tag{6}$$

In this way, the pixels with a similar pattern will be highlighted in the same activation map. And then, the foreground part-aware prototypes $\widetilde{\mathbf{P}} = \{\widetilde{\mathbf{p}}_k\}_{k=1}^{K}$ can be got by:

$$\widetilde{\mathbf{p}}_k = \text{GAP}(\mathbf{A}_\text{i} \odot \widetilde{\mathbf{F}}). \tag{7}$$

Substituting $\widehat{\mathbf{Y}}'$ with $(1 - \widehat{\mathbf{Y}}')$, we can get the background part-aware prototypes $\widetilde{\mathbf{P}} = \{\widetilde{\mathbf{p}}_k\}_{k=K+1}^{2K}$ in the same manner.

3.4 Training Objectives

During training, we calculate the supervise loss with the provided label $\mathbf{Y}_i^S$ of the source domain by:

$$\mathcal{L}_{sup} = \frac{1}{N} \sum_{i=1}^{N} \text{CE}(\widetilde{\mathbf{Y}}_i^S, \mathbf{Y}_i^S), \tag{8}$$

where the $\text{CE}(\cdot, \cdot)$ refers to the standard cross entropy loss.

Considering the part-aware prototypes may focus on the same, making the part miner degeneration, we impose a diversity loss to expand the discrepancy among part-aware prototypes. Formally,

$$\mathcal{L}_{div} = \sum_{i=1}^{2K} \sum_{j=1, i \neq j}^{2K} \cos(\widetilde{\mathbf{p}}_i, \widetilde{\mathbf{p}}_j), \tag{9}$$

where the $\cos(\cdot, \cdot)$ denotes cosine similarity between two vectors.

The discriminator $\mathcal{D}_{part}$ takes $\widetilde{\mathbf{p}}_k$ as input and outputs a scalar representing the probability that it belongs to the target domain. The loss function of $\mathcal{D}_{part}$ can be formulated as:

$$\mathcal{L}_{part} = \frac{1}{2K} \sum_{k=1}^{2K} [\text{CE}(\mathcal{D}_{part}(\widetilde{\mathbf{p}}_k^S), 0) + \text{CE}(\mathcal{D}_{part}(\widetilde{\mathbf{p}}_k^T), 1)]. \tag{10}$$

Table 1. Comparison with other SOTA methods on the Lucchi dataset. Note that "VNC III → Lucchi-Test" 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	VNC III → Lucchi-Test				VNC III → Lucchi-Train			
	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
Y-Net [19]	-	68.2	-	52.1	-	71.8	-	56.4
DANN [2]	-	68.2	-	51.9	-	74.9	-	60.1
AdaptSegNet [24]	-	69.9	-	54.0	-	79.0	-	65.5
UALR [29]	80.2	72.5	71.2	57.0	87.2	78.8	77.7	65.2
DAMT-Net [18]	-	74.7	-	60.0	-	81.3	-	68.7
DA-VSN [4]	82.8	75.2	73.9	60.3	91.3	83.1	82.2	71.1
DA-ISC [5]	89.5	81.3	80.5	68.7	92.4	85.2	84.5	74.3
Ours	**91.1**	**84.1**	**83.5**	**72.8**	**94.4**	**86.7**	**86.1**	**77.1**

Table 2. Comparison with other SOTA methods on the MitoEM dataset. Note that "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.

Method	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
UALR [29]	90.7	83.8	83.2	72.2	92.6	86.3	85.5	75.9
DAMT-Net [18]	92.1	84.4	83.7	73.0	94.8	86.0	85.7	75.4
DA-VSN [4]	91.6	83.3	82.6	71.4	94.5	86.7	86.3	76.5
DA-ISC [5]	92.6	85.6	84.9	74.8	96.8	88.5	88.3	79.4
Ours	**93.9**	**86.1**	**85.5**	**75.6**	**97.0**	**89.2**	**88.8**	**80.6**

As a result, the overall objective of our SAPAN is as follows:

$$\mathcal{L} = \mathcal{L}_{sup} + \lambda_{div} \times \mathcal{L}_{div} - \lambda_{part} \times \mathcal{L}_{part}, \tag{11}$$

where the λ_{div} and λ_{part} are the trade-off weights. The segmentation network and the $\mathcal{D}_{part}$ are trained alternately by minimizing the $\mathcal{L}$ and the $\mathcal{L}_{part}$, respectively.

4 Experiments

4.1 Dataset and Evaluation Metric

Dataset. We evaluate our approach on three widely used EM datasets: the VNC III [3] dataset, the Lucchi dataset [9] and the MitoEM dataset [28]. These datasets exhibit significant diversity making the domain adaptation task challenging. The VNC III [3] dataset contains 20 sections of size 1024×1024. The training and testing set of Lucchi [9] dataset are both $165 \times 1024 \times 768$. The MitoEM dataset consists of two subsets of size $1000 \times 4096 \times 4096$, dubbed MitoEM-R and MitoEM-H respectively. The ground-truth of their training set (400) and validation set (100) are publicly available.

Metrics. Following [5,29], four widely used metrics are used for evaluation, *i.e.*, mean Average Precision (mAP), F1 score, Mattews Correlation Coefficient (MCC) [14] and Intersection over Union (IoU).

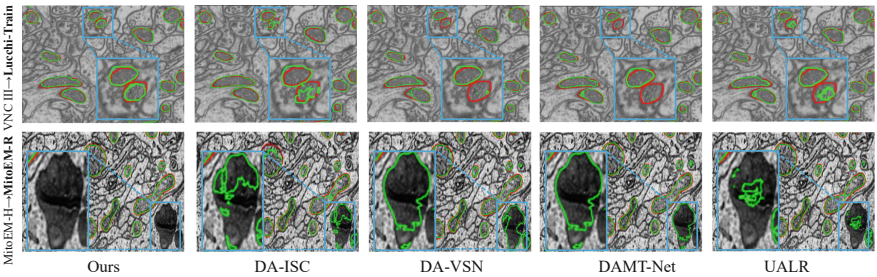

Fig. 2. Qualitative comparison with different methods. Note that the red and green contours denote the ground-truth and prediction. And we mark significant improvements using blue boxes. (Color figure online)

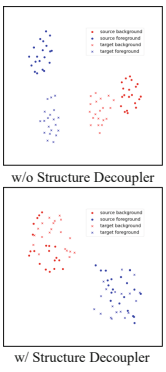

Fig. 3. t-SNE visualization.

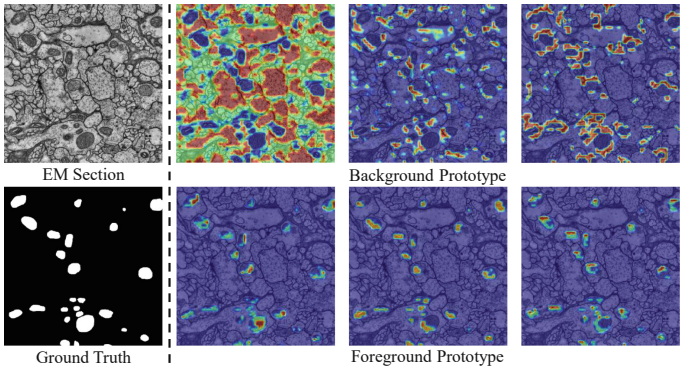

Fig. 4. Visualization of activation maps **A** for different prototypes in part miner.

Table 3. Ablation study on components.

SD	PM	mAP	F1	MCC	IoU
		85.3	80.3	79.4	67.5
✓		88.8	82.6	81.9	70.5
	✓	87.9	81.6	80.9	69.1
✓	✓	**91.1**	**84.1**	**83.5**	**72.8**

Table 4. Ablation study on smoothing operations (SMT).

	mAP	F1	MCC	IoU
w/o SMT	84.6	81.5	80.7	68.9
Conv. SMT	85.9	82.6	82.0	70.6
Atten. SMT	**91.1**	**84.1**	**83.5**	**72.8**

Table 5. Ablation on different ways for modeling part prototypes.

$fg.$	$bg.$	$\mathcal{L}_{div}$	mAP	F1	MCC	IoU
✓			75.3	75.5	74.7	61.2
✓		✓	88.4	82.2	81.4	69.9
✓	✓	✓	**91.1**	**84.1**	**83.5**	**72.8**

4.2 Implementation Details

We adopt a five-stage U-Net with feature channel number of $[16, 32, 48, 64, 80]$. During training, we randomly crop the original EM section into 512×512 with random augmentation including flip, transpose, rotate, resize and elastic transformation. All models are trained 20,000 iterations using Adam optimizer with batch size of 12, learning rate of 0.001 and β of (0.9, 0.99). The λ_{div}, λ_{part} and K are set as 0.01, 0.001 and 4, respectively.

4.3 Comparisons with SOTA Methods

Two groups of experiments are conducted to verify the effectiveness of our method. Note that "Oracle" means directly training the model on the target domain dataset with the ground truth and "NoAdapt" means training the model only using the source domain dataset without the target domain dataset.

Table 1 and Table 2 present the performance comparison between our method and other competitive methods. Due to the similarity between time series (video) and space series (EM volume), the SOTA domain adaptive video segmentation method [4] is also compared. We consistently observe that our method outperforms all other models on two groups of experiments, which strongly proves the effectiveness of our method. For example, our SAPAN enhances the IoU of VNC III → Lucchi-Test and Lucchi-Train to 72.8% and 77.1%, outperforming DA-ISC by significant margins of 4.1% and 2.8%. Compared with the pixel-wise alignment, the part-aware alignment can make the model fully utilize the global context information so as to perceive the mitochondria comprehensively.

On the MitoEM dataset with a larger structure discrepancy, our SAPAN can achieve 75.6% and 80.6% IoU on the two subsets respectively, outperforming DA-ISC by 0.8% and 1.2%. It demonstrates the remarkable generalization capacity of SAPAN, credited to the structure decoupler, which can effectively alleviate the domain gap caused by huge structural difference.

Figure 2 shows the qualitative comparison between our SAPAN and other competitive methods including UALR [29], DAMT-Net [18], DA-VSN [4] and ISC [5]. We can observe that other methods tend to incorrectly segment the background region or fail to activate all the mitochondria. We deem the main reason is that the large domain gap severely confuses the segmentation model. With the assistance of the structure decoupler and the part miner, SAPAN is

more robust in the face of the large domain gap and generates a more accurate prediction.

4.4 Ablation Study and Analysis

To look deeper into our method, we perform a series of ablation studies on VNC III → Lucchi-Test to analyze each component of our SAPAN, including the **S**tructure **D**ecoupler (SD) and the **P**art **M**iner (PM). Note that we remove all modules except the U-Net and the pixel-wise discriminator as our baseline. Hyperparameters are discussed in the supplementary material.

Effectiveness of Components. As shown in Table 3, both structure decoupler and part miner bring a certain performance lift compared with the baseline. (1) With the utilization of SD, a 3.0% improvement of IoU can be observed, indicating that decoupling the feature from domain-specific information can benefit the domain adaptation task. In Fig. 3, we visualize the feature distribution with/without SD using t-SNE [11]. We can see that SD makes the feature distribution more compact and effectively alleviates the domain gap. (2) The introduction of PM achieves further accuracy gains, mainly ascribed to the adaptive part alignment mechanism. As shown in Fig. 4, the different prototypes focus on significant distinct areas. The discriminator benefits from the diverse parts-aware prototypes, which in turn promotes the segmentation network.

Effectiveness of the Attention-Based Smoothing. As shown in Table 4, abandoning the spatial smoothing operation makes the performance decrease. Compared with simply employing a convolution layer for smoothing, attention-based smoothing contributes to a remarkable performance (72.8% *vs.* 70.6% IoU), thanks to the long-range modeling capabilities of the attention mechanism.

Effectiveness of the Ways Modeling Part-Aware Prototypes. In Table 5, $fg.$ means only focusing on foreground and vice versa. Neglecting $\mathcal{L}_{div}$ leads to severe performance degradation, that is because $\mathcal{L}_{div}$ is able to prevent the prototypes from focusing on similar local semantic clues. And simultaneously modeling foreground/background prototypes brings further improvement, demonstrating there is a lot of discriminative information hidden in the background region.

5 Conclusion

We propose a structure decoupler to decouple the distribution and morphology, and a part miner to aggregate diverse parts for UDA mitochondria segmentation. Experiments show the effectiveness.

Acknowledgments. This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62021001).

References

1. Duchen, M.: Mitochondria and Ca2+ in cell physiology and pathophysiology. Cell Calcium **28**(5–6), 339–348 (2000)
2. Ganin, Y., et al.: Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1) (2016). 2096-2030
3. Gerhard, S., Funke, J., Martel, J., Cardona, A., Fetter, R.: Segmented anisotropic sstem dataset of neural tissue. Figshare (2013)
4. Guan, D., Huang, J., Xiao, A., Lu, S.: Domain adaptive video segmentation via temporal consistency regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8053–8064 (2021)
5. Huang, W., Liu, X., Cheng, Z., Zhang, Y., Xiong, Z.: Domain adaptive mitochondria segmentation via enforcing inter-section consistency. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 89–98. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16440-8_9
6. Li, M., Chen, C., Liu, X., Huang, W., Zhang, Y., Xiong, Z.: Advanced deep networks for 3D mitochondria instance segmentation. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2022)
7. Li, Z., Chen, X., Zhao, J., Xiong, Z.: Contrastive learning for mitochondria segmentation. In: 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 3496–3500. IEEE (2021)
8. Liu, D., et al.: PDAM: a panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. IEEE Trans. Med. Imaging **40**(1), 154–165 (2020)
9. Lucchi, A., Li, Y., Fua, P.: Learning for structured prediction using approximate subgradient descent with working sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1987–1994 (2013)
10. Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Camouflaged instance segmentation via explicit de-camouflaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17918–17927 (2023)
11. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(11) (2008)
12. Mai, H., Sun, R., Zhang, T., Xiong, Z., Wu, F.: DualRel: semi-supervised mitochondria segmentation from a prototype perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19617–19626 (2023)
13. Martin, L.J.: Biology of mitochondria in neurodegenerative diseases. Prog. Mol. Biol. Transl. Sci. **107**, 355–415 (2012)
14. Matthews, B.W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure **405**(2), 442–451 (1975)
15. Newsholme, P., Gaudel, C., Krause, M.: Mitochondria and diabetes. An intriguing pathogenetic role. Adv. Mitochondrial Med., 235–247 (2012)
16. Nightingale, L., de Folter, J., Spiers, H., Strange, A., Collinson, L.M., Jones, M.L.: Automatic instance segmentation of mitochondria in electron microscopy data. BioRxiv, 2021–05 (2021)
17. Pan, Y., et al.: Adaptive template transformer for mitochondria segmentation in electron microscopy images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)

18. Peng, J., Yi, J., Yuan, Z.: Unsupervised mitochondria segmentation in EM images via domain adaptive multi-task learning. IEEE J. Sel. Top. Sig. Process. **14**(6), 1199–1209 (2020)
19. Roels, J., Hennies, J., Saeys, Y., Philips, W., Kreshuk, A.: Domain adaptive segmentation in volume electron microscopy imaging. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1519–1522. IEEE (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-aware transformers for diabetic retinopathy grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10938–10947 (2021)
22. Sun, R., et al.: Appearance prompt vision transformer for connectome reconstruction. In: IJCAI (2023)
23. Sun, R., Wang, Y., Mai, H., Zhang, T., Wu, F.: Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
24. Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481 (2018)
25. Wang, L., Tong, Z., Ji, B., Wu, G.: TDN: temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1895–1904 (2021)
26. Wang, Y., Sun, R., Zhang, T.: Rethinking the correlation in few-shot segmentation: a buoys view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7183–7192 (2023)
27. Wang, Y., Sun, R., Zhang, Z., Zhang, T.: Adaptive agent transformer for few-shot segmentation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) European Conference on Computer Vision, pp. 36–52. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19818-2_3
28. Wei, D., et al.: MitoEM dataset: large-scale 3D mitochondria instance segmentation from EM images. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12265, pp. 66–76. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59722-1_7
29. Wu, S., Chen, C., Xiong, Z., Chen, X., Sun, X.: Uncertainty-aware label rectification for domain adaptive mitochondria segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 191–200. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_18
30. Yi, J., Yuan, Z., Peng, J.: Adversarial-prediction guided multi-task adaptation for semantic segmentation of electron microscopy images. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1205–1208. IEEE (2020)