Unsupervised Mitochondria Segmentation in EM Images via Domain Adaptive Multi-Task Learning

Jialin Peng, Jiajin Yi, and Zhimin Yuan

***Abstract*—Semantic segmentation of mitochondria is essential for electron microscopy image analysis. Despite the great success achieved using supervised learning, it requires a large amount of expensive per-pixel annotations. Recent studies have proposed to exploit similar but annotated domains by domain adaptation, but the possible severe domain shift poses a challenge for the model transfer. In this study, we develop an unsupervised domain adaptation method to adapt the model trained on an labeled source domain to the unlabeled target domain. Specifically, we achieve cross-domain segmentation by integrating geometrical cues provided by the annotated labels and the visual cues latent in images of both domains in a framework of *adversarial domain adaptive multi-task learning*. Rather than enforcing manually-defined shape priors, we propose to learn geometrical cues from the source domain through adversarial learning. Domain-invariant and discriminative features are learned through joint adaptation. Extensive ablations, parameter analysis and comparisons have been conducted on three benchmarks under various settings. The experiments show that our method performs favorably against state-of-the-art methods both in segmentation accuracy and visual quality.**

***Index Terms*—Adversarial learning, electron microscopy, mitochondria segmentation, unsupervised domain adaptation.**

I. Introduction

THE accurate segmentation of electron microscopy (EM) images is an essential step for understanding the neuronal structures of a brain and its functions. Current state-of-the-art segmentation methods for medical images are supervised trained convolutional neural networks, across which a common obstacle is the dependence on large amounts of pixel-wise labeled data. The popular *encoder-decoder* architecture [1], [2], e.g., U-Net, has demonstrated effectiveness and versatility on many biomedical image segmentation tasks. However, annotating biomedical images including EM images is not only labor-intensive and time-consuming, but also need adequate expertise, which make it difficult to obtain a large number of labeled images. Instead of re-annotating in each domain, generalizing the model trained on the dataset with enough annotated labels (*source domain*) well to a novel dataset without labels (*target domain*) is an appealing alternative approach to mitigate the burden of manual labeling and the difficulties of training on unlabeled dataset. We will refer to this as the unsupervised scenario (as shown in Fig. 1), which is a difficult situation and usually addressed by unsupervised domain adaptation (UDA). Currently, UDA has usually been exploited to reduce the discrepancy between different domains. The challenge for cross-domain adaptation of EM segmentation models is the severe domain shift. Specifically, the set of EM images with annotations and the set without annotations may be not only scanned using different types of electron microscope but also taken from different biological tissues or even different species, which result in the large domain shift. An example is demonstrated in Fig. 1, where the images of the two domains are taken from the CA1 hippocampus region of a mouse brain, and the Drosophila melanogaster third instar larva Ventral Nerve Cord (VNC), respectively. Predictably, the size and density of mitochondria as well as the appearance of background tissues show obvious differences. Although the source and target tasks share almost the same label space (but with natural label shift), cross-domain adaptation still suffers from *significant shift* between the source and target domains.

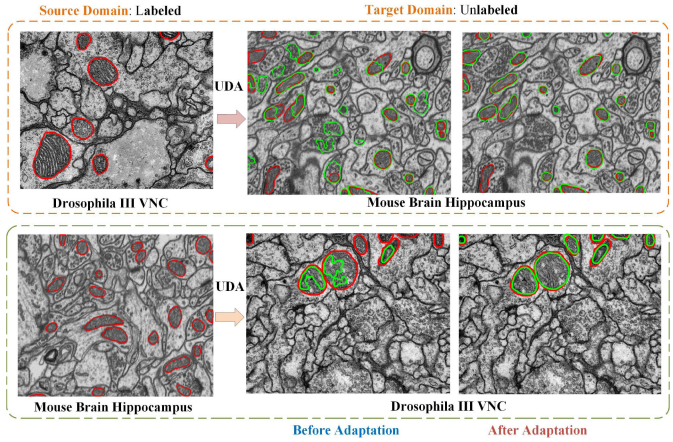

Fig. 1. Unsupervised Domain Adaptation (UDA) for the segmentation of tiny subcellular organelle from EM images. The image stack of Mouse Hippocampus is scanned by FIB-SEM, whereas Drosophila III VNC data are scanned by ssTEM. The source and target data are taken from different tissues of different species. The red contours correspond to the ground-truth, and the green contours correspond to the predictions by the model without adaptation (left) and the proposed DAMT-Net (right).

Moreover, semantic segmentation of tiny subcellular organelle (e.g., mitochondria) from EM images is a challenging

Manuscript received December 14, 2019; revised April 15, 2020, June 8, 2020, June 19, 2020, and June 19, 2020; accepted June 22, 2020. Date of publication June 26, 2020; date of current version September 24, 2020. This work was supported in part by NSFC under Grant 11771160, in part by the Science and Technology Program of Fujian (2019H0016), and in part by Huaqiao Univerisity under Grant ZQN-PY411. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Arrate Muñoz-Barrutia. *(Corresponding author: Jialin Peng.)*

The authors are with the College of Computer Science and Technology, Xiamen Key Laboratory of Computer Vision and Pattern Recognition, and the Fujian Key Lab of Big Data Intelligence and Security, Huaqiao University, Xiamen 361021, China (e-mail: 2004pjl@163.com; 839236375@qq.com; zhimin_yuan@163.com).

Digital Object Identifier 10.1109/JSTSP.2020.3005317

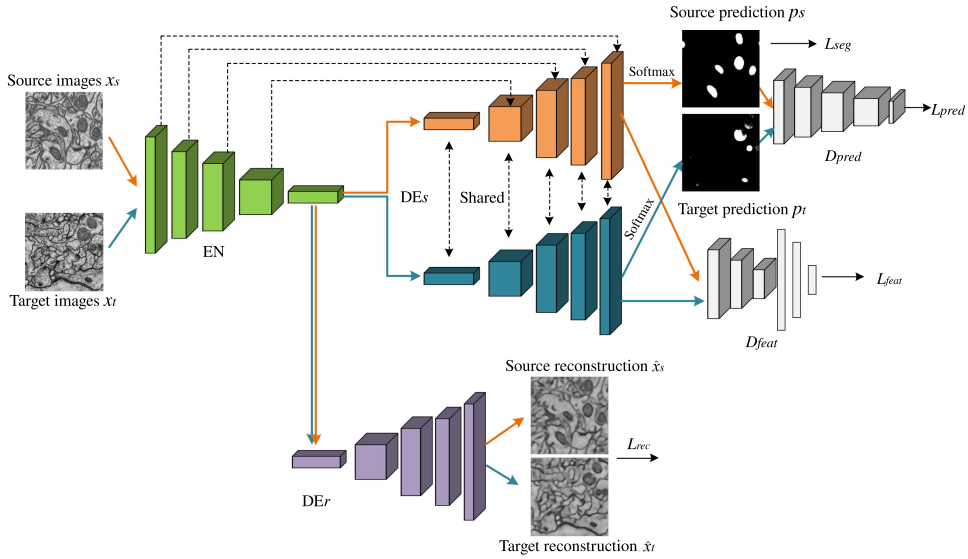

Fig. 2. Architecture of the proposed *Domain Adaptive Multi-Task Network* (DAMT-Net). We use the notable encoder-decoder network U-Net [1] as the backbone of segmentation network. An auto-encoder (in purple) is utilized as the image reconstruction network for domain-invariant visual feature learning. Since the learned features in the *encoder* (including those copied to *decoder*) are less discriminative, dual domain discriminators applied on the output-prediction space and output-feature space separately at the ending stage of the *decoder* to integrate geometrical cues and maintain discriminability.

task in itself, even in fully *supervised scenario* [3]. In fact, the mitochondria are unevenly and sparsely distributed, and show large variability in density, size and shape. Moreover, many sub-structures of a cell share similar intensities in EM images with mitochondria; strong contrasts do not necessarily correspond to the semantic boundaries of mitochondria. For the more challenging unsupervised cross-domain segmentation considered in this study, it is crucial to integrate high-level information (e.g., shape cues) provided by the annotated source domain and low-level visual cues latent in both the source and target images. Rather than enforcing shape prior (e.g., roundness) defined according to human knowledge, we propose to learn generalized geometrical cues from label-prediction space of source domain through adversarial learning.

Many UDA algorithms, particularly those for training deep networks, have been introduced for classification and segmentation problems in computer vision [4]–[12] and medical image analysis [13]–[17]. Whereas UDA has shown competitive performance on the classification task, its application on segmentation task is still challenging. One dominant strategy towards UDA seeks to learn domain-invariant features through feature distribution alignment using statistical distance metrics [4]–[6], [13], denoising auto-encoder [8], [14], domain-adversarial learning [9]–[11], etc. The alignment may happen at different representation layers of a deep network, whereas the prominent approaches seek to align feature representations at the end of encoding stage [8], [10], [11]. Compared with classification tasks, the pixel-level segmentation tasks are much more challenging due to the need of high-dimensional features to encode diverse visual cues, and carefully designed *decoder* is also crucial to recover fine segmentations. Therefore, only conducting distribution alignment at feature encoding stage and learning decoder only using source domain as [8], [10], [11] may not preserve discriminability very well for the target domain. Although domain-discriminator at output-space has been introduced in [9], the learned features and cross-domain predictor lack guidance form visual cues in the target image space, which are the most reliable information available from the unlabeled target domain.

To address these above problems, we propose a domain adaptive multi-task network, named DAMT-Net (as shown in Fig. 2), for unsupervised domain adaptive segmentation of EM images. The notable encoder-decoder network U-Net is employed as the backbone for segmentation generator. In line of much of previous work, we seek to learn domain-invariant representations whereas preserving discriminability. For this objective, we propose to exploit both geometrical cues and visual cues in the adversarial multi-task learning framework. Specifically, we learn domain-invariant visual features though auxiliary guidance from the reconstruction of both the source and target data. To utilize geometrical cues and improve the discriminative ability of the representation for segmenting the unlabeled target domain, we further guide the representation learning by adversarial learning in the semantic prediction space and feature space of the late decoding stage. In this way, the adaptation is jointly guided by both label domain and image information, and the distribution alignment is enforced on both the encoder and decoder. Compared with the methods in [8], [14] that only adapt features for image reconstruction, the proposed method adapts to learn domain invariant features that are more specific to the discriminative tasks. Extensive experiments on three datasets under various settings have shown the effectiveness of the proposed UDA method.

This work is an extensive extension of our preliminary work [18] with validations on more challenging datasets. More extensive experiments on performance analysis and parameter analysis as well as more visual analysis are conducted. More

comprehensive review of literature and detailed discussion of our motivation have also been included.

II. Related Work

In this section, we briefly discuss semantic segmentation methods and UDA for classification and segmentation.

Semantic segmentation: It is the task of associating each pixel of an image with an object label. State-of-the-art methods are mainly based on the recent advances of fully convolutional network [1], [19], e.g., FCN [19], Deeplab [2], which usually rely on pre-trained VGG [20] or ResNet101 [21] as backbones. The *encoder-decoder* networks [1], [2], e.g., U-Net, have demonstrated their effectiveness and versatility on many biomedical image segmentation tasks. Whereas the encoder performs multi-level feature representation learning, the decoder gradually recovers the spatial information and generates a semantic segmentation mask. In this study, we choose U-Net as our backbone network for segmentation.

UDA with deep networks: Recently, deep networks have proven effective at increasing the performance of domain adaptation. UDA using deep networks can be realized through various strategies. The dominant strategy is to learn a latent space with shared encoder, where the feature distributions of the source and target images can be aligned. Early studies usually achieve this though optimizing for some measurements of the distributional discrepancy such as maximum mean discrepancy (MMD) and its kernel variants [4]–[6]. Ghifary *et al.* [8] proposed to learn the domain-invariant feature representation with a deep reconstruction-classification network (DRCN), in which they augmented the supervised classification task with an auxiliary task to reconstruct images of the target domain using denoising auto-encoder. The similar idea was applied for EM image segmentation in [14] under the name Y-Net.

Another representative way of distribution matching is to train a domain discriminator [10], [11], [22] through adversarial training in a certain feature space. In the domain adversarial neural network (DANN) [10], a domain discriminator was employed on the output features of the *encoder* to ensure that, using the learned features, a classifier cannot distinguish between the source and target domain examples. Meanwhile, a discriminative classifier was trained with labels in the source domain, such that the learned representation is discriminative for source domain. Instead of utilizing shared *encoder* for domain-invariant feature learning as [8], [10], the adversarial discriminative domain adaptation method [11] used an asymmetric adaptation allowing independent source and target encoders, where the target encoder was learned with domain adversarial training as DANN.

However, these methods that conduct distribution alignment at feature encoding stages using MMD, auto-encoder, or domain discriminator may not preserve much discriminability on target domain for the pixel-wise segmentation problem. Compared with classification tasks, the pixel-level segmentation tasks are much challenging due to the need of encoding diverse visual cues, and carefully designed decoder is crucial to recover spatial information and fine segmentations [1]. Rather than learning the *decoder* using only the source domain data as DANN, DRCN and Y-Net, we propose to improve the task-specific discriminability of the learned features through reducing discrepancy at both encoding and decoding stages.

Recently, Bermúdez-Chacón *et al.* [13] designed a two-stream U-Net for semi-supervised domain-adaptive segmentation of EM images. Specifically they proposed to use two partially different but related segmentation generators for the source and target domains separately, and they minimized a MMD loss to reduce the feature discrepancy at the final stage of the *decoders*. Rozantsev *et al.* [23] extended the method in [13] to a generalized two-stream architecture for both classification and segmentation tasks, where the weights in all corresponding layers of the two streams are related but not shared. One limitation of two-stream networks [13], [23] is the doubled parameters for untying layers, which results in unwanted memory requirement. In [9], [12], output-level adaptation with adversarial training has been exploited by assuming spatial similarities between the semantic labeling spaces of the two domains, and sound performance has been obtained on semantic segmentation of cityscapes scenarios [9]. In this work, we follow the similar idea to align representations at decoding stage and learn geometrical cues in label space. Note that the label space of the multi-label problems in cityscapes scenarios contains much more rich contexts than our binary task of EM segmentation. Therefore, we further incorporate a domain discriminator on features at the ending stage of the *decoder*. Other recent studies [7], [24], [25] directly transform the source images to resemble the target images with generative models, which is ready to be utilized to enhance other domain adaptation methods [25].

III. Method

Let the source domain $D_S = \{X_S, Y_S\}$ be a set of labeled images drawn from a source domain distribution $p_s(x, y)$, and the target domain $D_T = \{X_T\}$ be a novel set of unlabeled images drawn from the target domain distribution $p_t(x, y)$. Distributions $p_s(x, y)$ and $p_t(x, y)$ are both assumed unknown and similar, but different $p_s(x, y) \neq p_t(x, y)$. We define the labeled sample of the source data as $x_s \in X_S$, and the corresponding label as $y_s \in Y_S$. Similarly, the sample of unlabeled target data is defined as $x_t \in X_T$. Our objective is to learn a label prediction model from the unlabeled target domain and labeled source domain such that it can produce sound segmentation on the target domain.

A. Overview of the Proposed Model

An overview of our proposed DAMT-Net is depicted in Fig. 2. It is a deep feed-forward encoder-decoder architecture with a shared encoder EN to aggregate multi-level features, and two separate decoders DE_s and DE_r for segmentation and image reconstruction tasks, respectively. Typically, by training the segmentation generator with the labeled source images, the DE_s is only discriminative for the source labels. For unsupervised domain adaptive segmentation, our goal is to learn a good cross-domain label predictor, in which the learned features, both in its encoder and decoder, are not only domain-invariant,

but also discriminative, especially on the target domain. To this end, we perform joint adaptation on two stages, namely encoding stage and decoding stage, in the paradigm of multi-task learning. Specifically, on encoding stage, we learn features that can reconstruct images of both domains following the idea of DRCN and Y-Net [8], [14]. On the decoding stage, domain-adversarial adaptations in both the label prediction space and outputting feature space are conducted to make the features be domain-invariant and discriminative for the target domain.

More specifically, we decompose the *encoder-decoder* into three parts. We assume that each input sample x from the source or target domain is first mapped into a feature representation by the encoder mapping EN. After that, the feature representation is mapped to the label space and original image space by segmentation decoder DE_s and reconstruction decoder DE_r, respectively. The *encoder* EN and *decoder* DE_s jointly define the segmentation generator, whereas the *encoder* EN and *decoder* DE_r jointly define an auto-encoder [26] as the data reconstruction pipeline. The two pipelines (tasks) share the same encoding parameters, which we want to be domain invariant and furthermore discriminative, especially on the target domain. Following the practices of prior works [8], [10], learning a shared feature mapping can significantly reduce the number of model parameters compared to using untied parameters [11], [23]. However, it usually only ensures that the learned feature representation is only discriminative on the source domain, but less discriminative on the target domain especially for the challenging semantic segmentation task. To address this problem, we employ adversarial adaptation on the late decoding stage of DE_s to guide the discriminative feature learning, which is explained in Section III-B.

B. The Model

We denote the generated features by the layer preceding the output of our DAMT-Net for the source data and target data as f_s and f_t, respectively. The label predictions for the source and target data are denoted by p_s and p_t, respectively.

Encoding stage adaptation: Recall that our goal is to minimize the domain discrepancy between representations of the source and target data. We propose to learn shared feature representation capable of encoding enough visual information from both domains. To this end, we augment an auto-encoder path EN+DE_r (the green and purple layers in Fig. 2) to the main segmentation generator path EN+DE_s, where the two paths share the same feature encoder EN. The encoding representations in EN are shared to encode invariant visual cues for the source data x_s and target data x_t. When feeding x_s and x_t to segmentation generator, the decoder of DE_s produces predictions p_s and p_t, and the DE_r branch produces the reconstructed images $\hat{x}_s$ and $\hat{x}_t$, respectively. It is expected that the reconstruction data $\hat{x}_s$ and $\hat{x}_t$ are close to the corresponding original data x_s and x_t respectively, whereas the prediction p_s of x_s is close to the corresponding label y_s. We use cross-entropy loss L_{seg} for the segmentation and mean squared loss L_{rec} for reconstruction as follows,

$$\mathcal{L}_{seg} = -\mathbb{E}_{(x_s,y_s)}\left[\sum_{c=1}^{2} y_s^c \log(p_s^c)\right], \tag{1}$$

$$\mathcal{L}_{rec} = \mathbb{E}_{x_s}\left[||x_s - \hat{x}_s||_2^2\right] + \mathbb{E}_{x_t}\left[||x_t - \hat{x}_t||_2^2\right], \tag{2}$$

where spatial dimensions are omitted for simplicity in Eq. (1).

Although the reconstruction branch is unsupervised and not for learning discriminative features, it aims to capture the visual information in the image space to support the adaptation to the target domain. Note that the auxiliary reconstruction strategy has been exploited in many tasks including domain adaptation [8], [14]. To improve the discriminability of the learned features for the segmentation task on the target domain, we further conduct adaptation on the decoding stage.

Decoding stage adaptation: We use shared segmentation generator for the two domains as illustrated in Fig. 2. Taking either the encoded representation of source images x_s or the target images x_t as input, we obtain the final feature maps f_s, f_t, that are the representations preceding the prediction layer, and label predictions p_s, p_t, respectively.

Ideally, the shared representation learned in the encoding adaptation stage is not only domain invariant and also maintaining label discriminative on both domains. However, generally this is not the case, especially on the target domain. The source segmentation model still shows degenerated discriminative ability on the target domain (shown in Table III) due to the unsupervised learning of domain-invariant representation and the natural shift in label space between different domains.

So far, we have used visual information in images from both domains to learn invariant representation, and used label information only from the source domain to learn discriminative features. However, we still lack the ability of discrimination on the target domain, and the decoding representations are still less biased to the target domain. To this end, we further guide the feature learning by adversarial learning on both prediction and feature spaces: 1) on the structured prediction space, we follow the idea of [9] to leverage domain label through adversarial learning; 2) on the space of the decoding representation, we enhance the discriminative ability by applying a domain discriminator on decoding features f_s and f_t. The adversarial learning on multiple spaces aims to ensure that the network cannot distinguish between the feature distributions of the source and target domains, and enable the encoding and decoding representations to be more discriminative on the unlabeled target domain.

More specifically, we minimize multi-level discrepancy of source and target representations through alternating minimization between two objective functions. First, we utilize two domain discriminators D_{pred} and D_{feat} separately to classify whether an image is drawn from the source or the target domain. The discriminator D_{pred} is applied on the final predictions (Fig. 2), whereas the D_{feat} is applied on the final decoding representations preceding the output layer (Fig. 2). In this way, the domain level supervision can be utilized; the two discriminators are trained to distinguish domain labels of inputs from different

domains. The discriminators D_{pred} and D_{feat} are learned by minimizing the following standard supervised loss, where the label 1 indicates the source domain and the label 0 indicates the target domain.

$$\min_{D_{pred},D_{feat}} \quad -\lambda_{feat}\mathcal{L}_{feat} - \lambda_{pred}\mathcal{L}_{pred}, \tag{3}$$

in which

$$\mathcal{L}_{pred} = \mathbb{E}_{x_s}[\log D_{pred}(p_s)] + \mathbb{E}_{x_t}[\log(1 - D_{pred}(p_t))], \tag{4}$$

$$\mathcal{L}_{feat} = \mathbb{E}_{x_s}[\log D_{feat}(f_s)] + \mathbb{E}_{x_t}[\log(1 - D_{feat}(f_t))], \tag{5}$$

in which the width and height dimensions are omitted for simplicity in Eq. (4).

Second, when the parameters of discriminator network are frozen, the segmentation-reconstruction network is learned through minimizing the following loss function.

$$\min_{EN,DE_s,DE_r} \quad \mathcal{L}_{seg} + \lambda_{rec}\mathcal{L}_{rec} - \lambda_{feat}\mathcal{L}^{'}_{feat} - \lambda_{pred}\mathcal{L}^{'}_{pred}, \tag{6}$$

in which λ_{rec}, λ_{feat} and λ_{pred} are trade-off weights and $\mathcal{L}^{'}_{feat}$ and $\mathcal{L}^{'}_{pred}$ are defined as following.

$$\mathcal{L}^{'}_{feat} = \mathbb{E}_{x_t}[\log(D_{feat}(f_t))]. \tag{7}$$

$$\mathcal{L}^{'}_{pred} = \mathbb{E}_{x_t}[\log(D_{pred}(p_t))]. \tag{8}$$

With inverted domain label, minimizing $-\mathcal{L}^{'}_{feat}$ and $-\mathcal{L}^{'}_{pred}$ corresponds to maximizing the probability of the target prediction being considered as the source one. As a result, it will enable the learned features to fool the discriminators D_{pred} and D_{feat}, which will improve the domain-invariant feature learning. During the testing time, only the segmentation generator EN+DE_s is kept and applied.

C. Network Architecture

Segmentation network EN+DE_s: Given the effectiveness of U-Net for segmentation, we adopt the architecture of U-Net [1] with minor modification, which is an *encoder-decoder* with skip connections, as depicted in Fig. 2. Compared with the original U-Net, 1) padding is used to make the output have equal size with the input; 2) bilinear interpolation is used for up-sampling; 3) group normalization [27] is utilized on each layer, since we rely on small batch size.

Reconstruction network EN+DE_r: We use *encoder-decoder* architecture as the auto-encoder for reconstruction of original images. Specifically, the *decoder* for reconstruction is the same as the *decoder* in U-Net but without skip connections between the *encoder* and *decoder*.

Feature discriminator D_{feat}: The architecture consists of 3 convolution layers with kernel size 3×3 and stride 1, followed by max pooling, and 3 fully-connected layers, where the channel number is $\{48, 48, 48, 288, 144, 2\}$. ReLU activation and group normalization [27] are used.

Prediction discriminator D_{pred}: For the discriminator applied on predictions, we use a 5-layer fully-convolutional neural network with group normalization and leaky ReLU parameterized by 0.2. Following [9], the kernel size is 4×4, and stride is 2, where the channel number is {64, 128, 256, 512, 1} and padding is 1. Leaky ReLU parameterized by 0.2 and group normalization [27] are used.

IV. Experimental Results

A. Data Description

We evaluate our DAMT-Net on the challenging task of unsupervised mitochondria segmentation on three benchmarks.

Mouse Hippocampus: The EPFL Mouse Hippocampus Data[1] [28] are taken from CA1 hippocampus region of a mouse brain and scanned by focused ion beam scanning electron microscope (FIB-SEM). It contains two subsets, each of which is of size $165 \times 1024 \times 768$ with pixel-wise label. The Subset 1 is typically used for model testing, where the Subset 2 is used for model training. The voxel resolution is approximately $5 \times 5 \times 5$ nm^3/voxel.

Drosophila III VNC: This image stack[2] [29] contains 20 sections from serial section Transmission Electron Microscopy (ssTEM) of the Drosophila melanogaster third instar larva VNC. This image stack is of size $20 \times 1024 \times 1024$, which measures $4.7 \times 4.7 \times 1$ microns approx., with a resolution of 4.6×4.6 nm^2/pixel and section thickness of 45-50 nm. Besides the different scanning instruments and different biological tissues, it also has a significant different voxel resolution from the image cube from Mouse Hippocampus.

Drosophila I VNC: This image stack[3] contains 30 sections of size 512×512 from a ssTEM dataset of the Drosophila first instar larva VNC. The image cube measures $2 \times 2 \times 1.5$ microns approx., with a resolution of $4 \times 4 \times 50$ nm^3/voxel.

We validate the proposed model under three settings: 1) Drosophila III VNC→Mouse Hippocampus, 2) Mouse Hippocampus→Drosophila III VNC and 3) Mouse Hippocampus→Drosophila I VNC. For each setting, we train the model on the source domain with supervision and test to adaptation to the target domain without any supervision.

B. Experimental Setup

Evaluation metrics: Our segmentation tasks are both binary segmentation tasks. Given a ground truth (Y) and a segmentation (X), we evaluate the segmentation accuracy by comparing the segmentation to manual segmentation using Dice similarity coefficient (DSC) and Jaccard index (JAC),

$$\mathrm{DSC}(\mathrm{X},\mathrm{Y}) = \frac{2 \times |\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X}| + |\mathrm{Y}|}, \tag{9}$$

$$\mathrm{JAC}(\mathrm{X},\mathrm{Y}) = \frac{|\mathrm{X} \cap \mathrm{Y}|}{|\mathrm{X} \cup \mathrm{Y}|}, \tag{10}$$

where $|\cdot|$ returns the number of pixels contained in a set.

For detection evaluation, F1 Score is used as the measure.

$$\text{F1 Score} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{11}$$

[1][Online]. Available: https://cvlab.epfl.ch/data/em

[2][Online]. Available: https://github.com/unidesigner/groundtruth-drosophila-vnc

[3][Online]. Available: https://www.ini.uzh.ch/ acardona/data.html

TABLE I
SEARCH SPACE AND BEST ASSIGNMENTS OF THE BASIC SETTING HYPER-PARAMETERS FOR ALL THE COMPARED ADAPTATION METHODS

Hyper-parameter	Search space	Best assignment
number of epochs	100	100
patience	10	10
batch size	1	1
optimizer	Adam	Adam
learning rate	$\{0.1, 1, 10\}\times10^{-4}$	1×10^{-4}
momentum β_1	0.9	0.9
momentum β_2	0.99	0.99
learning rate scheduler	polynomial decay (0.9)	polynomial decay (0.9)

TABLE II
THE SETTINGS FOR THE TRADE-OFF HYPER-PARAMETERS

Methods	λ_{rec} ($\times10^{-3}$)	λ_{feat} ($\times10^{-3}$)	λ_{pred} ($\times10^{-3}$)	λ^1_{pred} ($\times10^{-3}$)	λ^1_{seg} ($\times1$)
Search space	$\{0, 0.1, 0.2, 0.5, 1, 1.5, 3, 5\}$				
NoAdapt	-	-	-	-	-
Y-Net [14]	1.0	-	-	-	-
AdaptSegNet [8]	-	-	1.0	0.2	0.1
DANN [9]	-	1.0	-	-	-
Our model	0.5	1.0	1.0	-	-

where TP, FP, and FN denote the true positives, false positives and false negatives, respectively. For both segmentation and detection tasks, the performance is computed slice by slice.

Implementation details: We implement our network using PyTorch on a 1080Ti GPU with 11 GB memory. For the training of the segmentation network (i.e., U-Net) on source domain, we use the Adam optimizer [30], where the momentum is set as 0.9 and 0.99. The initial learning rate is set as 5×10^{-5} and is decreased using polynomial decay with a power of 0.9. For our proposed network, we use the Adam optimizer with the learning rate as 1×10^{-4} and polynomial decay with a power of 0.9. For the model training, the trade-off hyper-parameters are set as λ_{rec}=5×10^{-4}, $\lambda_{feat} = \lambda_{pred} = 1 \times 10^{-3}$. As suggested by [31], we have reported the search space and best assignments of hyper-parameters for our method and the compared methods in Tables I and II.

To train large networks from limited training data and prevent overfitting, we apply a large variety of data augmentation techniques on the fly during training: 1) random rotations, flip and transpose, 2) random intensity shift and scale, 3) random elastic deformations, 4) motion blur, and 5) random crop. At the inference time, we apply test-time augmentation including 3 variations of flips. The modified U-Net is trained with the input patches of size 512×512 and batch size 3. Our network is trained using input patches of the same size and batch size 1. The training process is stopped if the loss does not improve after 10 epochs. The model training takes about 8 hours. The code and pre-trained model are available at https://github.com/Jiajin-Yi/DAMT-Net.

Baselines methods: Although UDA for semantic segmentation has been intensively explored, quite few studies [13], [14] have directly investigated the problem of EM image segmentation. Thus, we have implemented several related state-of-the-art methods [9], [10] for unsupervised domain adaptation in natural computer vision. Specifically, we compare our method with the following representative approaches that can operate without annotations in the target domain:

a) **NoAdapt** model is trained without consideration of the target-domain data. Specifically, it directly applies the modified U-Net trained on the labeled source domain to segment images of the target domain.
b) **Y-Net** [14] seeks to simultaneously reconstruct images of both domain and train a segmentation generator, which is an extension of DRCN model [8] to the segmentation task.
c) **DANN** [10] applies a domain discriminator on the output of the encoder shared by the source and target domains, aiming to make the two domains be indistinguishable in feature space.
d) **AdaptSegNet** [9] aligns feature distributions through adversarial domain adaptation in output-label space. *Multi-level* output-label spaces [9] are employed to enhance adaptation. We also evaluate the AdaptSegNet approach using *single-level* output-label space for adaptation.
e) **Supervised method** uses labeled data from the target domain to train the modified U-Net. This is an upper bound of the our unsupervised domain adaptive segmentation method.

For a fair comparison, we use the same U-Net as the segmentation generator for all the compared methods. All the methods including ours are trained using randomly cropped patches of size 512×512 and batch size 1; the same training augmentation and testing augmentation as that in our method are applied. For the competing methods such as Y-Net and AdaptSegNet, we use the trade-off parameters suggested in their papers/codes by default. As for the weight of the feature discriminator in DANN, we use the same value as the weight of our feature discriminator D_{feat} by default. The parameter settings for all the methods are listed in Table II. Since all the compared methods are unsupervised model, we use the training subset on the target domain as the validation set. For hyper-parameter selection on validation set, we have also suggested the search range. Note that λ^1_{pred} and λ^1_{seg} are the hyper-parameters involved in the multi-scale adaptation in [9].

C. Drosophila III VNC→Mouse Hippocampus

To evaluate the proposed model, we first perform experiments on the EPFL Mouse Hippocampus data. It is the most widely-used benchmark for evaluation of mitochondria segmentation. This dataset is a large dataset containing two subset, i.e., Subset 1 and Subset 2, each of which is of size $165 \times 1024 \times 768$. For evaluation, we use 2-fold cross validation, that is using one subset as the training data of the unlabeled target domain and the other subset as the testing data. The Drosophila III VNC dataset of size $20 \times 1024 \times 1024$ is used as the labeled source domain. The two datasets are scanned with different electron microscopes. Fig. 1 depicts the concept of this cross-domain unsupervised segmentation task with example images. As for observing the obviously different appearances of the target object as well as the backgrounds, there is severe domain shift.

TABLE III
COMPARISON WITH STATE-OF-THE-ART UDA METHODS FOR MITOCHONDRIA SEGMENTATION ON MOUSE HIPPOCAMPUS DATASET, WHICH CONTAINS TWO SUBSETS NAMED SUBSET 1 AND SUBSET 2. TWO-FOLD CROSS VALIDATION IS USED FOR EVALUATION. THE ASTERISK INDICATES STATISTICALLY SIGNIFICANT DIFFERENCE ($p < 0.05$) COMPARED WITH OUR METHOD

Methods	Drosophila III VNC→Mouse Hippocampus			
	Subset 1		Subset 2	
	DSC (%)	JAC (%)	DSC (%)	JAC (%)
Source domain (supervised)	88.0	78.7	88.0	78.7
NoAdapt	57.3*	40.3*	61.3*	44.3*
Y-Net [14]	68.2*	52.1*	71.8*	56.4*
AdaptSegNet [8](single-level)	69.2*	53.1*	77.8*	64.0*
AdaptSegNet [8](multi-level)	69.9*	54.0*	79.0*	65.5*
DANN [9]	68.2*	51.9*	74.9*	60.1*
Our model	74.7	60.0	81.3	68.7
Supervised (Full Target)	92.7*	86.5*	93.9*	88.6*
Supervised (10% Target)	88.7*	80.0*	91.3*	84.0*

TABLE IV
ABLATION STUDY OF OUR JOINT METHOD. DE_r: ENCODING STAGE ADAPTATION BY AUTO-ENCODER; DE_{feat}: DECODING STAGE ADAPTATION BY ALIGNING FEATURES; DE_{pred}: DECODING STAGE ADAPTATION BY ALIGNING PREDICTIONS

Experiments	Drosophila III VNC→Mouse Hippocampus			
	Subset 1		Subset 2	
	DSC (%)	JAC (%)	DSC (%)	JAC (%)
NoAdapt	57.3*	40.3*	61.3*	44.3*
DE_r	68.2*	52.1*	71.8*	56.4*
DE_{feat}	69.2*	53.1*	77.2*	63.1*
DE_{pred}	69.2*	53.1*	77.8*	64.0*
DE_r+DE_{feat}	72.1*	56.5*	79.4*	66.1*
DE_r+DE_{pred}	74.1*	59.1*	79.0*	65.5*
$DE_{feat+pred}$	71.2*	55.5*	78.9*	65.4*
$DE_{pred+pred}$	69.9*	54.0*	79.0*	65.5*
Our full model	74.7	60.0	81.3	68.7

Segmentation performance on Mouse Hippocampus: Table III summarizes the adaptation results in comparison with NoAdapt and state-of-the-art UDA methods [9], [10], [14]. Due to the domain shift, when applying the supervised trained model, which has achieved a score of 88.0% in DSC on the source domain, to the target domain, it can only obtain a score of 57.3% in DSC on the Subset 1 and 61.3% on the Subset 2. Note that the DSC score of NoAdapt on the Subset 2 is 4.0% higher than that on the Subset 1 due to the uneven distribution of mitochondria. All of the compared UDA methods are able to significantly improve the performance over NoAdapt. Especially, the proposed method performs favorably against other UDA algorithms on both subsets, and improves the DSC over the NoAdapt method by 17.4% on the Subset 1 and 20.0% in DSC on the Subset 2.

Whereas the competing UDA methods use different feature layers for adaptation, the results in Table III shows the benefit of the joint adaptation with visual cues and geometrical cues by our methods. Especially, compared to the cutting edge method AdaptSegNet which conducts multi-level adaptation in output label space, our method also shows better accuracy with a performance gain in DSC of 4.8% on the Subset 1 and 2.3% on the Subset 2. The reason is that we further constrain the domain-invariant feature learning process by directly modeling the visual cues, which are the most crucial and reliable information for imaging data of the target domain. However, with only visual cues, the Y-Net has only achieved a DSC score of 68.2% on the Subset 1 and 71.8% on the Subset 2, which are 6.5% and 9.5% lower than our method on the two subset, respectively. Compared with the DANN model, we use the same adversarial domain discriminator (i.e., DE_{feat}) in feature space but on the late representation layer in the decoder. With only DE_{feat}, our model can still achieve a score of 69.2% in DSC and 53.1% in JAC on the Subset 1 (see Table IV), whereas the DANN model achieves a lower score of 68.2% in DSC and 51.9% in JAC. With additional visual and geometrical cues, our full model achieves a performance gain of 6.5% in DSC and 8.1% in JAC. Furthermore, *t*-test has indicated the statistical significance of the difference between our method and other methods.

In addition, another factor to evaluate the adaptation performance is to measure the gap between the adaptation model and the supervised trained models with different amount of labeled data on the target domain, which are also listed in Table III. Predictably, there is still a large gap between the unsupervised methods and fully supervised segmentation methods. However, with 20% of the training set of the Mouse Hippocampus data, i.e., a volume of $33 \times 1024 \times 768$, the gap measured by DSC between our method and the supervised model narrows down to -14.0% on the Subset 1 and to -10.0% on the Subset 2. Besides, note that the data augmentation in training stage can be essential for supervised model to achieve sound performance with scarce labeled training data.

Visual comparison on Mouse Hippocampus: Fig. 3 shows typical segmentation results of the unsupervised cross-domain adaptation from Drosophila III VNC to Mouse Hippocampus. Each column in Fig. 3 presents one typical example, from up to down: (a) the ground truth; (b) segmentations without adaptation; (c) results of Y-Net; (d) segmentation of AdaptSegNet; (d) results of DANN; (f) results by our model. In Fig. 3, orange boxes are used to highlight some challenging regions. Although the mitochondria in all the three examples show varying number, distribution, size as well as shape, our method visually shows superior segmentation accuracy and more regular shape.

Parameter analysis using the Subset 1: We have tested the effectiveness of the trade-off hyper-parameters, i.e., λ_{rec}, λ_{feat}, and λ_{pred} on the Subset 1 of Mouse Hippocampus dataset. As shown in experiment setup section, we set $\lambda_{feat} = \lambda_{pred}$, since the feature space and prediction space carry information in similar level. Thus, we first tune the λ_{rec} in the range of $\{0, 0.1, 0.5, 1, 1.5, 3, 5\} \times 10^{-3}$, while fixing $\lambda_{feat} = \lambda_{pred}$ as the default value 1×10^{-3}. Then, we tune $\lambda_{feat} = \lambda_{pred}$, when fixing λ_{rec} as 0.5×10^{-3}. The results are depicted in Fig. 4. It can be seen that our method is relatively stable in the range of $[0.1, 3] \times 10^{-3}$. However, tuning the parameter can still boost the performance. As shown in Fig. 4, with $\lambda_{feat} = \lambda_{pred} = 1.5 \times 10^{-3}$ and $\lambda_{rec} = 0.5 \times 10^{-3}$, the proposed model can achieve the best performance (i.e., 75.7% in DSC and 61.2%

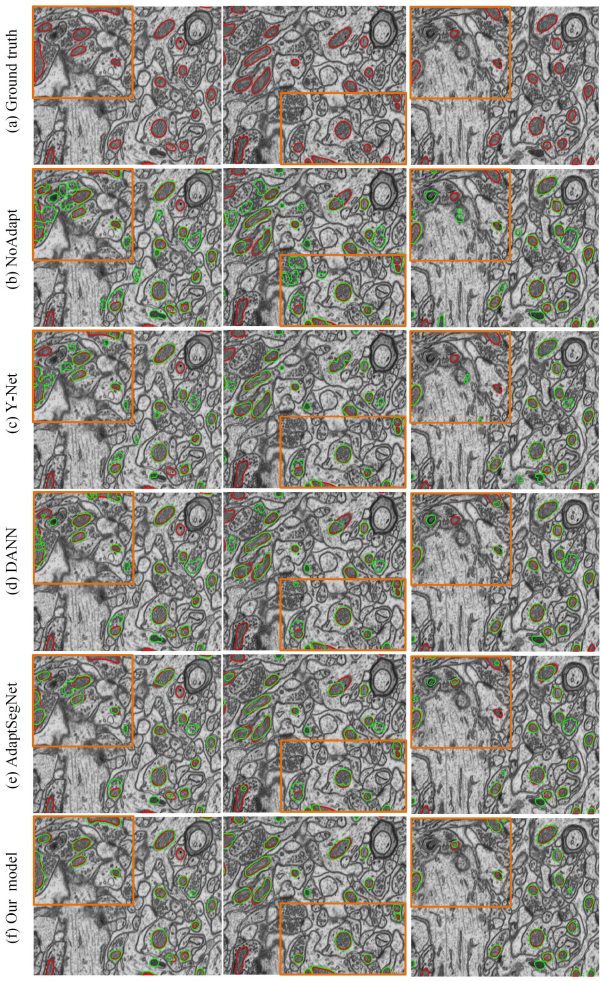

Fig. 3. Example results of the unsupervised domain-adaptive segmentation of Mouse Hippocampus data. Each column presents one example slice, where both predictions and ground truth are overlaid. The green contours correspond to the predictions of different methods, and the red contours to the ground-truth. Orange boxes are used to highlight some region of interest.

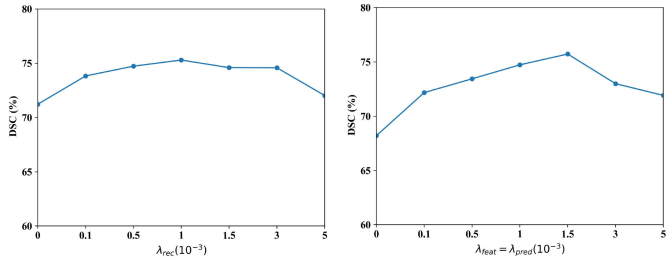

Fig. 4. Impact of the tradeoff hyper-parameters on the performance for Drosophila III VNC→Mouse Hippocampus adaptation (Subset 1). (a) The impact of tuning λ_{rec}, when fixing $\lambda_{feat} = \lambda_{pred}$ as 1×10^{-3}; (b) the impact of tuning $\lambda_{feat} = \lambda_{pred}$, when fixing λ_{rec} as 0.5×10^{-3}. Our default settings are $\lambda_{rec} = 0.5 \times 10^{-3}$ and $\lambda_{feat} = \lambda_{pred} = 1 \times 10^{-3}$.

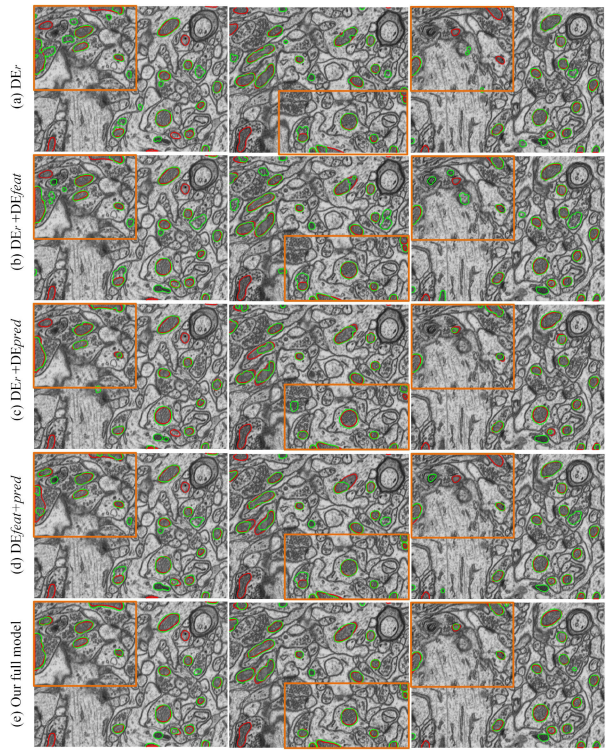

Fig. 5. Visual comparison of the ablated versions of our model for Drosophila III VNC→ Mouse Hippocampus (Subset 1). The green contours correspond to the predictions of different methods, and the red to the ground-truth.

in JAC), which is 1.0% in DSC and 1.2% in JAC higher than that using the default parameters.

Ablation study on Mouse Hippocampus: To investigate the effectiveness of the different modules of our joint encoding and decoding adaptation approach, seven ablated versions of our model have been compared: 1) DE_r (Y-Net): only encoding-stage adaptation with auto-encoder; 2) DE_{feat}: only adaptation in the feature space of the late DEcoding stage; 3) DE_{pred}: only adaptation in the prediction space; 4) DE_r+DE_{feat}: DE_r with additional adaptation in the feature space of the late DEcoding stage; 5) DE_r+DE_{pred}: DE_r with additional adaptation in the prediction space; 6) $DE_{feat+pred}$: the joint adaptation in both feature and prediction spaces; 7) $DE_{pred+pred}$ (AdaptSegNet): the joint adaptation in multi-level prediction spaces; 8) our full model (DAMT-Net): DE_r+$DE_{feat+pred}$.

The quantitative comparison results on the two subsets of Mouse Hippocampus dataset are shown in Table IV, and the visual comparison results are depicted in Fig. 5. In comparison with the baseline method (NoAdapt), it is obvious that the method with the DE_r outperforms the baseline method by a large margin, which indicates that the cross-domain prediction model is more biased towards the target domain. Moreover, as shown in Table IV the performance of DE_{feat} and DE_{pred} respectively is superior than the baseline method and DE_r, which confirms the effectiveness of the feature space and prediction space adaptation. Moreover, the combination of DE_r with DE_{feat} as well as DE_{pred} obviously perform better than using them separately. Visually, by adapting with additional geometrical cues in prediction space, the results of DE_r+DE_{pred} show more regular shape as shown in Fig. 5 (a) and (c). By integrating multiple cues with DE_r+$DE_{feat+pred}$, our full model achieves the best performance. Visually, our model also shows obviously reduced false positive and false negative detections as illustrated

TABLE V
COMPARISON WITH STATE-OF-THE-ART UDA METHODS ON DROSOPHILA III VNC DATASET, WHICH IS SPATIALLY DIVIDED INTO TWO SUBSETS FOR TWO-FOLD CROSS VALIDATION

Methods	Mouse Hippocampus→Drosophila III VNC			
	Subset 1		Subset 2	
	DSC (%)	JAC (%)	DSC (%)	JAC (%)
Source domain (supervised)	92.7	86.5	92.7	86.5
NoAdapt	37.1*	23.6*	45.4*	30.5*
Y-Net [14]	71.0*	55.2*	66.1*	49.7*
AdaptSegNet [8](single-level)	73.3*	58.1*	68.3*	52.4*
AdaptSegNet [8](multi-level)	73.6*	58.6*	69.0*	53.2*
DANN [9]	72.1*	56.7*	67.2*	51.1*
Our model	77.0	62.8	71.7	56.6
Supervised model (Full Target)	88.7*	79.9*	87.2*	77.5*
Supervised model (20% Target)	81.5*	69.0*	79.9*	67.2*

in Fig. 5(e), and the shape of the detected mitochondria is also more regular.

All above experimental results show that the adaptation jointly guided by geometrical cues in label space and image visual cues is an effective way to mitigate the domain gap and improve segmentation accuracy.

D. Mouse Hippocampus→Drosophila III VNC

We also have conducted experiments on the adaptation from Mouse Hippocampus dataset to Drosophila III VNC dataset. Since the Drosophila III VNC dataset is a relatively small dataset of size $20 \times 1024 \times 1024$, we split the data volume along the x axis into 2 subsets with equal size ($20 \times 512 \times 1024$), and use 2-fold cross validation for model evaluation. Note that due to the sparsely and unevenly distributed mitochondria, one subset may be more challenging than the other one.

Segmentation performance on Drosophila III VNC: Table V summarizes the semantic segmentation performance of the adaptation from Mouse Hippocampus to Drosophila III VNC. Similar to the results on Mouse Hippocampus dataset, our method yields the best performance in terms of both DSC and JAC on the two subsets. Specifically, we outperform the Y-Net by 6.0% and 5.6% in DSC on the two subsets respectively, and surpass the multi-level AdaptSegNet by 3.4% and 2.7% in DSC on the two subsets respectively. In addition, the gaps between of our unsupervised method and fully supervised method on the two subsets are -11.7% and -15.5% in DSC, respectively. When training a supervised model with reduced amount of labeled training data (i.e., 20%), the gaps between of our unsupervised method and fully supervised method on the two subsets reduce to -4.5% and -8.2%, respectively.

Visual comparison on Drosophila III VNC: Fig. 6 presents several example results for the unsupervised adaptive segmentation. The first two columns and last two columns are examples from different subsets. As shown in Fig. 6, the mitochondria in the four example images show obviously varying appearances, sizes and shapes. Yellow dashed boxes indicate mitochondria successfully detected by our method but falsely detected by other methods; orange dashed boxes are used to highlight the

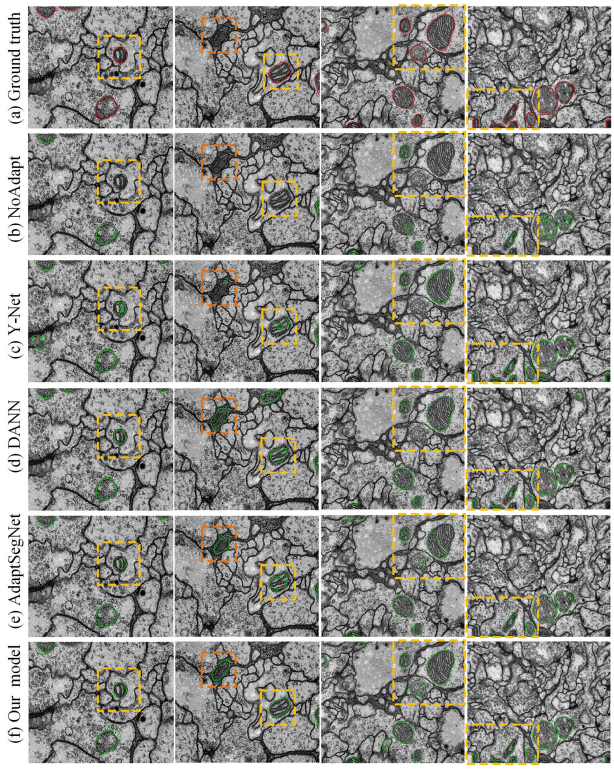

Fig. 6. Example results of the unsupervised domain adaptive segmentation for Mouse Hippocampus→Drosophila III VNC. The green contours correspond to the predictions of different methods, and the red to the ground-truth.

TABLE VI
COMPARISON WITH STATE-OF-THE-ART UDA METHODS FOR THE ADAPTATION FROM MOUSE HIPPOCAMPUS TO DROSOPHILA I VNC

Methods	Mouse Hippocampus→Drosophila I VNC	
	DSC (%)	JAC (%)
Source domain (supervised)	92.7	86.5
NoAdapt	46.5*	32.2*
Y-Net [14]	61.9*	46.6*
AdaptSegNet [8](single-level)	68.7*	54.1
AdaptSegNet [8](multi-level)	67.9*	53.2
DANN [9]	62.5*	46.9*
Our model	71.5	57.2
Supervised (Drosophila III VNC)	77.4*	67.1*

mitochondria falsely detected by our method. Compared to other methods, our method visually shows better segmentation accuracy.

Taking the image in the 1st column of Fig. 6 for example, although the mitochondrion in yellow box show significant different appearance with the mitochondria in the source domain images, our method can segment it with accuracy. Another challenging case is the image in the 3rd column. Specifically, the three mitochondria in the yellow box show quite different sizes and textures. With no adaptation, NoAdapt model fails to detect two of the three mitochondria, and other three state-of-the-art methods also fail to detect part of them due to the distinct appearances. In contrast, our method can successfully detect all of them. There is also a failure case shown in the 2nd column of Fig. 6, our method as well as most of other methods segment

TABLE VII
PERFORMANCE COMPARISON ON VALIDATION SETS. SINCE ALL THE COMPARED METHODS ARE UNSUPERVISED, IN THE 2-FOLD CROSS VALIDATION WE USE THE TRAINING SUBSET ON THE TARGET DOMAIN AS THE VALIDATION SET

Methods	Drosophila III VNC→Mouse Hippocampus				Mouse Hippocampus→Drosophila III VNC				Mouse Hippocampus→Drosophila I VNC	
	Subset 1		Subset 2		Subset 1		Subset 2		Drosophila III VNC (Subset1)	
	DSC (%)	JAC (%)	DSC (%)	JAC (%)	DSC (%)	JAC (%)	DSC (%)	JAC (%)	DSC (%)	JAC (%)
NoAdapt	57.3*	40.3*	61.3*	44.3*	37.1*	23.6*	45.4*	30.5*	37.1*	23.6*
Y-Net [14]	64.8*	48.1*	70.2*	54.3*	57.4*	41.3*	66.4*	50.7*	57.4*	41.3*
AdaptSegNet [8]	74.9*	59.9*	74.4*	59.6*	72.7*	57.5*	66.7*	51.3*	72.7*	57.5*
DANN [9]	70.5*	54.5*	74.3*	59.3*	69.0*	53.1*	66.1*	50.4*	69.0*	53.1*
Our model	77.4	63.1	78.6	64.9	77.5	63.4	67.6	52.3	77.5	63.4

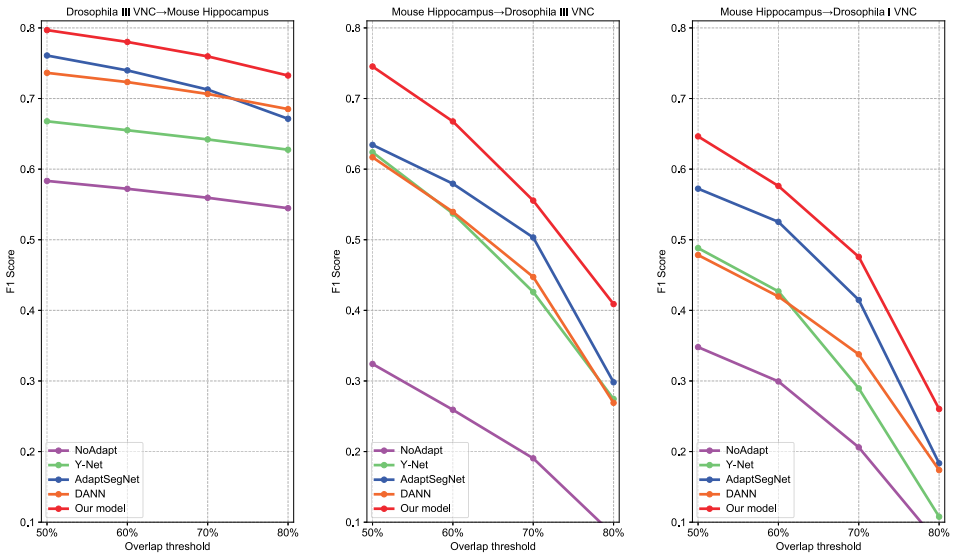

Fig. 7. Average detection performance of our method under different overlap thresholds. The mitochondria detection performance is computed slice by slice.

the non-mitochondrion object (in orange box) as mitochondrion. Last but not least, visually the segmented objects in our results show more regularized shape than methods without considering label space similarity (e.g., NoAdapt, Y-Net, and DANN), which may additionally indicate the effectiveness of utilizing geometrical cues.

E. Mouse Hippocampus→Drosophila I VNC

We have also conducted experiment on the adaptation from Mouse Hippocampus data to Drosophila I VNC dataset, which is a small dataset. Since the Drosophila I VNC dataset and Drosophila III VNC dataset are taken from similar biological tissues and both acquired with ssTEM (but with different resolutions). We use the whole Drosophila I VNC dataset as unlabeled testing dataset in the target domain, the Drosophila III VNC dataset (Subset1) as the unlabeled training data in the target domain, and use Mouse Hippocampus dataset as the labeled source domain. Accordingly, the Supervised method used as upper bound is trained on Drosophila III VNC dataset (subset1). The results are summarized in Table VI. Again, our model consistently outperforms other methods, and the gap between our method and the Supervised is -5.9% in DSC.

F. Segmentation Performance on Validation Sets

Since our method is unsupervised and uses no label information on the target domain, in Table VII we also report the segmentation results on the validation sets in the 2-fold cross validation in previous sections. Similar to the results on testing sets, our method consistently show superior results.

G. Detection Performance

Besides pixel-wise segmentation performance, we have also assessed the mitochondria detection performance. Since the Drosophila III VNC, and I VNC datasets have very large inter-slice thickness and have very small number of mitochondria, we evaluated the detection performance slice-by-slice to show the ability of our method to detect tiny objects. Average performance across the 2-fold cross validation is computed. The comparative results on all of the three EM datasets are shown in Fig. 7. For performance evaluation, we first conduct connected component analysis and then count the number of TP, FP and FN. TP detections are positive-predicted connected components that have T% (e.g., 50%) or more overlap with that in the ground truth, while FP detections are positive-predicted connected components that overlap with the ground truth mitochondria below T%. FN detections are ground truth mitochondria that do not

intersect with the predicted mitochondria. Finally, the F1 Score is computed according to Eq. (11). Compared to other baseline methods, our method has achieved the best detection accuracy under various overlap thresholds (T%), which further confirms the promising performance of our method.

V. Conclusion

In this study, we have proposed a multi-task adaptation network to address the problem of unsupervised cross-domain semantic segmentation of EM images. We showed that integrating of label space information, decoding feature information and image visual cues can improve the discriminative ability of the cross-domain label predictor on unlabeled target domain. More specifically, for domain-adaptive segmentation, label space information can provide high-level geometrical information (e.g., shape, size, etc), whereas information used for image reconstruction can provide valuable low-level cues. The method has been validated on unsupervised domain adaptation tasks with severe domain shifts. Experimental results showed that our proposed UDA method can achieve state-of-the-art performance for mitochondria segmentation from EM images.

Despite the improved results, the performance of domain adaptation methods for unsupervised cross-domain segmentation with large domain shift is still quite limited. In the future work, we will exploit structured visual information through self-learning on target domain to improve domain adaptation. It is based on the observation that, even with our joint adaptation method, few knowledge about the target domain has been exploited.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[2] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 801–818.

[3] J. Peng and Z. Yuan, "Mitochondria segmentation from em images via hierarchical structured contextual forest," *IEEE J. Biomed. Health Informat.*, doi: 10.1109/JBHI.2019.2961792.

[4] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[5] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 443–450.

[6] M. Long, Y. Cao, and J. Wang, "Learning transferable features with deep adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[7] J. Hoffman *et al.*, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.

[8] M. Ghifary, W. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 597–613.

[9] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7472–7481.

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, and H. Larochelle, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.

[11] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7167–7176.

[12] Y. Tsai, K. Sohn, S. Schulter, and M. Chandraker, "Domain adaptation for structured output via discriminative patch representations," in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 1457–1465.

[13] R. Bermúdez-Chacón, P. Márquez-Neila, M. Salzmann, and P. Fua, "A domain-adaptive two-stream u-net for electron microscopy image segmentation," in *Proc. Int. Symp. Biomed. Imag.*, 2018, pp. 400–404.

[14] J. Roels, J. Hennies, Y. Saeys, W. Philips, and A. Kreshuk, "Domain adaptive segmentation in volume electron microscopy imaging," in *Proc. Int. Symp. Biomed. Imag.*, 2019, pp. 1519–1522.

[15] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 691–697.

[16] J. Ren, I. Hacihaliloglu, E. Singer, D. J. Foran, and X. Qi, "Adversarial domain adaptation for classification of prostate histopathology whole-slide images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2018, pp. 201–209.

[17] W. Yan, Y. Wang, M. Xia, and Q. Tao, "Edge-guided output adaptor: Highly efficient adaptation module for cross-vendor medical image segmentation," *IEEE Signal Process. Lett.*, vol. 26, no. 11, pp. 1593–1597, Nov. 2019.

[18] J. Yi, Z. Yuan, and J. Peng, "Adversarial-prediction guided multi-task adaptation for semantic segmentation of electron microscopy images," in *Proc. Int. Symp. Biomed. Imag.*, 2020, pp. 1205–1208.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3431–3440.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[22] J. Hoffman, D. Wang, F. Yu, and T. Darrell, "FCNS in the wild: Pixel-level adversarial and constraint-based adaptation," 2016, *arXiv:1612.02649*.

[23] A. Rozantsev, M. Salzmann, and P. Fua, "Beyond sharing weights for deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 801–814, Apr. 2019.

[24] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 3722–3731.

[25] L. Tran, K. Sohn, X. Yu, X. Liu, and M. Chandraker, "Gotta adapt'em all: Joint pixel and feature-level domain adaptation for recognition in the wild," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2672–2681.

[26] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[27] Y. Wu and K. He, "Group normalization," in *Proc. Eur. Conf. Computer Vision*, 2018, pp. 3–19.

[28] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua, "Supervoxel-based segmentation of mitochondria in em image stacks with learned shape feature," *IEEE Trans. Med. Imag.*, vol. 31, no. 2, pp. 474–486, Feb. 2012.

[29] S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter, "Segmented anisotropic ssTEM dataset of neural tissue," figshare. Dataset. 2013. [Online]. Available: https://doi.org/10.6084/m9.figshare.856713.v1

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[31] J. Dodge, S. Gururangan, D. Card, R. Schwartz, and N. A. Smith, "Show your work: Improved reporting of experimental results," in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 2185–2194.