3D Domain Adaptive Instance Segmentation via Cyclic Segmentation GANs

Leander Lauenburg, Zudi Lin, Ruihan Zhang, Márcia dos Santos, Siyu Huang, Ignacio Arganda-Carreras, Edward S. Boyd, Hanspeter Pfister, Fellow, IEEE, and Donglai Wei

Abstract—3D instance segmentation for unlabeled imaging modalities is a challenging but essential task as collecting expert annotation can be expensive and time-consuming. Existing works segment a new modality by either deploying pre-trained models optimized on diverse training data or sequentially conducting image translation and segmentation with two relatively independent networks. In this work, we propose a novel Cyclic Segmentation Generative Adversarial Network (CySGAN) that conducts image translation and instance segmentation simultaneously using a unified network with weight sharing. Since the image translation layer can be removed at inference time, our proposed model does not introduce additional computational cost upon a standard segmentation model. For optimizing CySGAN, besides the CycleGAN losses for image translation and supervised losses for the annotated source domain, we also utilize self-supervised and segmentation-based adversarial objectives to enhance the model performance by leveraging unlabeled target domain images. We benchmark our approach on the task of 3D neuronal nuclei segmentation with annotated electron microscopy (EM) images and unlabeled expansion microscopy (ExM) data. The proposed CySGAN outperforms pre-trained generalist models, feature-level domain adaptation models, and the baselines that conduct image translation and segmentation sequentially. Our implementation and the newly collected, densely annotated ExM zebrafish brain nuclei dataset, named NucExM, are publicly available at https://connectomics-bazaar.github.io/proj/CySGAN/index.html.

Index Terms—3D Instance Segmentation, Unsupervised Domain Adaptation, Expansion Microscopy (ExM), Electron Microscopy (EM), Zebrafish, Neuronal Nuclei.

I. INTRODUCTION

THE 3D Instance segmentation of cell nuclei is an essential topic attracting both biomedical and computer vision researchers [1]–[5]. Supervised deep learning with in-domain annotations (e.g., U-Net [6], [7]) has become the dominant methodology for mainstream imaging modalities. However, such an approach is less applicable for novel imaging modalities, e.g., expansion microscopy (ExM) [8], due to the lack of existing labels and the high annotation costs for newly collected data. This work focuses on segmenting a new imaging modality without any in-domain annotation (Fig. 1a).

Two common approaches try to overcome the challenges by leveraging existing labels from mainstream domains. One approach is to train a supervised model on diverse datasets (i.e., a generalist model) and apply it directly to the new domain [3], [4]. However, when the domain gap becomes too large, generalist models can produce unsatisfactory predictions without in-domain finetuning that requires new training labels. The other approach, known as unsupervised domain adaptation, usually involves unpaired image-to-image translation models like CycleGAN [9] and segments a new domain with a two-stage pipeline. The first stage translates the source images to match the target domain distribution, aiming to be indistinguishable from the target images while keeping the source structures. The second stage pairs the translated images and corresponding ground-truth labels in the source domain to train a supervised model. The optimized model can then segment real images in the target domain\(^1\) (Fig. 1b).

\(^1\)Expansion microscopy [8] alleviates the resolution limitation in optical microscopy by physically expanding the tissues.

\(^2\)The opposite way, which transfers the target domain images to the source domain and applies a supervised model trained on the source data, is also reasonable. However, the community uses it less often as this direction requires both the translation and segmentation models at inference time.
Fig. 1. Overview of the task and methods. (a) We aim to segment 3D instances in a completely unlabeled target domain ($I_Y$) by leveraging the images ($I_X$) and masks ($S_X$) in the source domain (i.e., unsupervised domain adaptation). Instead of (b) conducting image translation (e.g., via CycleGAN [9]) and instance segmentation as two separate steps, we propose (c) Cyclic Segmentation GAN (CySGAN) to unify the two functionalities using weight sharing, which is optimized with both image translation as well as supervised and semi-supervised segmentation losses.

Contributions We present CySGAN, a novel 3D domain adaptive instance segmentation method that segments instances in an unlabeled domain using a multi-task network. We introduce an augmentation-restoration cycle-consistency strategy that significantly enhances CySGAN’s spatial awareness and robustness without disrupting the generator–discriminator balance. Furthermore, we contribute a new densely annotated ExM zebrafish brain nucleus dataset, NucExM, as well as the training and inference code, to the research community.

II. RELATED WORKS

A. Unpaired Image-to-Image Translation

In biomedical domains, paired images from different imaging modalities are usually expensive or even infeasible to obtain. Therefore, unpaired image-to-image translation [9], [16] based on Generative Adversarial Networks (GAN) [17] becomes a sensible methodology to transfer source images to the target distribution. An exemplary framework usually consists of a generator that maps the source images to the target domain and a discriminator that decides whether an input image is from the real target distribution or synthesized. The generator is optimized with the gradients of the GAN loss back-propagated through the discriminator. CycleGAN [9] achieves impressive performance by ensuring cycle consistency when transferring translated images back to the source domain using a pair of symmetric generators. Further improvements include shared high-level layers [18] and latent space alignment [10]. We refer readers to the survey by Pang...
et al. [19] for a more detailed discussion of image-to-image translation literature. Specifically, our work combines image translation with segmentation models to tackle unlabeled modalities, extending a standard 3D segmentation with one additional output channel optimized with image translation objectives to adapt to the target distributions. Our proposed CySGAN simplifies existing frameworks that conduct image translation and segmentation using two separate networks.

B. Instance Segmentation of 3D Microscopy

3D instance segmentation from microscopy images is challenging due to the dense distribution of objects and unavoidable physical limitations in imaging (e.g., data is frequently anisotropic with uneven resolution among different axes). Recent learning-based approaches tackle these challenges by first optimizing CNN-based models to predict representations calculated from the instance masks, including object boundary [6], [20], [21], affinity map [15], [22], star-convex distance [4], flow-field [3] and the combination of multiple representations [5]. Watershed transform [23], [24] and graph partition [25] can then be applied to convert the predicted representations into instance masks. However, most existing works train the segmentation models in a supervised learning manner using in-domain annotations, which becomes infeasible considering the cost of acquiring expert annotations for new modalities. Our work focuses on unifying segmentation approaches with image translation to segment instances in new domains via unsupervised domain adaptation. At inference time, the image-translation component of CySGAN can be removed, which means CySGAN does not increase the deployment cost upon a standard 3D segmentation model.

C. Domain Adaptive Segmentation

We focus on unsupervised domain adaptation with unlabeled target data. Existing approaches can be categorized into appearance-level and feature-level adaptation methods.

For appearance-level adaptation, utilizing unsupervised image translation is a practical methodology. Chartsias et al. [26] designed a two-stage framework that first translates source images to the unlabeled domain using CycleGAN [9] and then trains a separate segmentation model using the synthesized images and source labels. However, since the two modules are optimized independently, the limited awareness of the translation network to the downstream segmentation task can restrict the performance. CyCADA [10], SIFA [13], EssNet [11] and SECGAN [12] improve the sequential model by jointly optimizing the translation and segmentation networks. However, using two separate networks increases the system complexity in training and deployment. The authors of CyCADA [10], for example, stated that although the model is theoretically end-to-end trainable, they need to train it in stages as it is too memory-intensive to optimize the full objective. Different from the mentioned works, we unify image translation and segmentation into a single model to significantly reduce the system complexity. Since the translation and segmentation layers base their predictions on the same high-level features, the CySGAN model enforces the consistency between translated images and segmentation maps from an architectural perspective.

Feature-level adaptation methods commonly optimize a model for two (or more) domains so that the outputs and high-level features from different domains are indistinguishable in distribution. For the unlabeled domain, adversarial losses are usually applied to enforce the alignment. For example, SIFA [13] uses GAN losses to minimize the gap between the segmentation predictions from the real and synthesized target-domain images. Tsai et al. [27] designs a model directly taking the source and target images as inputs and applying adversarial losses to align the high-level feature maps. Following existing works, we implement a feature-level adaptation model for 3D instance segmentation and show that our CySGAN and appearance-level adaptation models can achieve significantly better performance in neuronal nuclei segmentation.

To our best knowledge, the only existing work that explores joint translation and segmentation with weight sharing is SUSAN [14], but our work differs from it in two main aspects. First, SUSAN and most works mentioned above are for 2D semantic segmentation, while our work focuses on the more challenging 3D instance segmentation. Second, SUSAN only applies supervised segmentation losses to the annotated domain, while our CySGAN leverages semi-supervised losses for the unlabeled domain in the absence of ground-truth labels.

III. Method

In this section, we first give an overview of the CySGAN framework (Sec. III-A). We then present the image translation (Sec. III-B) and segmentation (Sec. III-C) objectives to optimize the system, as well as our implementation (Sec. III-D).

A. The CySGAN Framework

Suppose we have an annotated source domain $X = (I_X, S_X)$ where $I_X$ and $S_X$ denote the images and paired segmentation labels, respectively. For an unlabeled target domain $Y$ with only images $I_Y$, the goal is to generate the instance segmentation $S_Y$ without acquiring any manual annotations in $Y$. One straightforward approach is to use some domain adaptation method $F$ to synthesize images $I'_Y = F(I_X)$ that are indistinguishable from the distribution of $I_Y$ but keep the instance structure in $S_X$. Then a supervised model can be optimized using $(I'_Y, S_X)$ pairs, which predicts $S_Y$ from $I'_Y$ at inference time (Fig. 1b).

Sequentially conducting the translation and segmentation suffers from multiple weaknesses. First, the translation model is not designed with an end task in mind and can propagate errors to the second step. Second, the translation model does not benefit from the powerful structural guidance that instance segmentation can impose upon it. Third, two separate modules make the system complicated in training and deployment. Thus, we propose a framework that shares weights between the translation and instance segmentation. Our framework uses two generators - one per domain - that output both translated images and segmentation simultaneously (Fig. 1c):

$$F: I_X \rightarrow (I'_Y, S_X) \quad G: I'_Y \rightarrow (I_X, S_Y)$$ (1)
We denote the proposed framework as the cyclic segmentation GAN (CySGAN). Specifically, for an image $x_1 \sim I_X$, we have $[\hat{y}_i, \hat{x}_s] = F(x_1)$, where $\hat{y}_i$ is the synthesized image, $\hat{x}_s$ contains the predicted instance representations, and $[\hat{y}_i, \hat{x}_s]$ is their concatenation along the channel dimension. For the clarity in the following formulations, we also denote $\hat{y}_i = F(x_1)[i]$ and $\hat{x}_s = F(x_1)[s]$. Note that $G(F(x_1))$ is no longer a valid expression as both models take only an image as input but output the translated image and segmentation.

Fig. 2 shows the architecture of our CySGAN framework. For the segmentation part, each of the two generators yields the three instance representations binary foreground mask (B), instance contour map (C), and signed distance transform (D) from which we derive the instance masks (detailed in Sec.III-C). Therefore, a single generator simultaneously outputs the synthesized image and the three instance representations as four different output channels. In particular, $\hat{y}_i = F(x_1)[i]$ has a single channel while $\hat{x}_s = F(x_1)[s]$ has three channels, but with the same spatial dimensions (the same for $G$). Unlike previous works that sequentially conduct image translation and segmentation, our design decreases the system complexity. Moreover, since the translation and segmentation modules base their predictions on the same high-level features in the generator networks, our model implicitly increases the structural consistency between synthesized images and predicted segmentation maps from an architectural perspective.

At inference time, only the generator $G$ is required to segment $I_Y$. Besides, the output layer for image translation can be simply removed without influencing the prediction of the segmentation maps. Therefore, our CySGAN model does not introduce any additional computational cost in deployment.

In the following parts, we discuss how to effectively optimize CySGAN with multiple objectives and data augmentations. Different from standard unsupervised image translation, the two domains are asymmetric, as $X$ is labeled, while $Y$ is unlabeled. We thus apply similar image translation losses but unique segmentation losses for $X$ and $Y$ domains.

**B. Image Translation Losses**

Given an input image $x_1 \sim I_X$, we can denote $F$ as the forward generator and $G$ as the backward generator (Eqn. 1). Since paired $I_X$ and $I_Y$ are difficult or even infeasible to obtain, $F$ is usually optimized using the adversarial loss so that the real and synthesized images gradually become indistinguishable in terms of distribution:

$$L_{GAN}(F, D_Y) = \log D_Y(y_i) + \log(1 - D_Y(\hat{y}_i))$$

where $D_Y$ is the $I_Y$ discriminator, while $y_i$ and $\hat{y}_i$ are true and synthesized images ($\hat{y}_i = F(x_1)[i]$), respectively. Following CycleGAN [9], we additionally use a backward generator $G$ and discriminator $D_X$ for $I_X$ to symmetrically optimize $L_{GAN}(G, D_X^F)$ for translating $I_Y$ to $I_X$, as well as enforcing the cycle-consistency loss for the images in both domains:

$$L_{cycle}(F, G) = \|G(\hat{y}_i) - x_i\|_1 + \|F(x_i) - y_i\|_1$$

The GAN and cyclic losses enable the models to transfer images between $I_X$ and $I_Y$ distributions. However, the training of the original binary cross-entropy GAN loss (Eqn. 2) can be unstable. Therefore, following the official CycleGAN implementation, we instead optimize the LSGAN [28] loss:

$$L_{LSGAN}(F, D_Y) = (D_Y(y_i) - 1)^2 + (D_Y(\hat{y}_i) + 1)^2$$

This loss formulation has been shown to prevent vanishing gradient and smooth the training process. A symmetric adversarial loss is applied to optimize $G$. In our proposed CySGAN, the image translation losses do not affect the output layers for the segmentation maps, but it does change the backbone shared by both translation and segmentation modules.

**C. Instance Segmentation Losses**

1) **Labeled Source Domain:** Instance segmentation approaches for microscopy images [3–5], [21] usually predict instance representations computed from the permutation-invariant labels and then apply a decoding algorithm to yield the masks. In this work, we follow U3D-BCD [5] that predicts...
the binary foreground mask (B), instance contour map (C), and signed distance transform (D) as three output channels using a 3D U-Net [7], which are decoded by a marker-controlled watershed (MW) algorithm. The B and C channels are optimized with the binary cross-entropy loss (BCE), while D is regressed with the mean squared error (MSE). Given an image-label pair \((x_i, x_s)\) sampled from \((X, S_X)\), the loss is

\[
L_{seg}(F) = L_{bce} \left( F(x_i)^B, x_s^B \right) + L_{bce} \left( F(x_i)^C, x_s^C \right) + \| F(x_i)^D - x_s^D \|_2^2
\]  

(5)

where \(x_s = [x_s^B, x_s^C, x_s^D] \) is the concatenation of the three representations. For the supervised direction, the segmentation loss \(L_{seg}(F)\) of the forward generator and segmentation loss \(L_{seg}(G)\) (based on the synthesized \(\hat{y}_i\) of the backward generator) are optimized by directly comparing \(\hat{x}_s\) and \(\hat{y}_s\) with \(x_s\) from \(S_X\) (1 and 2 in Fig. 3a).

The loss \(L_{seg}(G)\) effectively trains \(G\) in a supervised manner to predict the segmentation representations. Moreover, this design is not restricted to a particular set of instance representations and can be easily modified to incorporate other methods\(^4\). In the next part, we present a set of novel losses to improve the quality of predictions in the absence of paired labels.

\(^4\)For example, SUSAN [14] applies the supervised segmentation losses for 2D semantic masks with pixel-wise class annotations.

\(\hat{x}_s\) and \(\hat{y}_s\) are optimized with the binary cross-entropy loss (BCE), while \(\hat{x}_s\) is regressed with the mean squared error (MSE). Given an image-label pair \((x_i, x_s)\) sampled from \((X, S_X)\), the loss is

\[
L_{seg}(F) = L_{bce} \left( F(x_i)^B, x_s^B \right) + L_{bce} \left( F(x_i)^C, x_s^C \right) + \| F(x_i)^D - x_s^D \|_2^2
\]  

(5)

where \(x_s = [x_s^B, x_s^C, x_s^D] \) is the concatenation of the three representations. For the supervised direction, the segmentation loss \(L_{seg}(F)\) of the forward generator and segmentation loss \(L_{seg}(G)\) (based on the synthesized \(\hat{y}_i\) of the backward generator) are optimized by directly comparing \(\hat{x}_s\) and \(\hat{y}_s\) with \(x_s\) from \(S_X\) (1 and 2 in Fig. 3a).

The loss \(L_{seg}(G)\) effectively trains \(G\) in a supervised manner to predict the segmentation representations. Moreover, this design is not restricted to a particular set of instance representations and can be easily modified to incorporate other methods\(^4\). In the next part, we present a set of novel losses to improve the quality of predictions in the absence of paired labels.

Our method is connected to semi-supervised learning as we incorporate unlabeled images in optimization using losses without paired labels. We can also choose other semi-supervised objectives, e.g., augmentation consistency [29], when the model takes images in the unlabeled domain as inputs. Our work emphasizes the concept of leveraging unlabeled images in a unified translation-segmentation framework, while the specific design choices can vary.

\section{D. Implementation}

\subsection{1) Full Objective:}
The full objective (\(L\)) of CySGAN is the sum of losses in Sec. III-B and III-C, which is

\[
L = L_{GAN}(F, D_X^F) + L_{GAN}(G, D_X^G) + L_{seg}(F, G)
\]

\[
+ L_{seg}(F) + L_{seg}(G)
\]

(7)

We assign a uniform weight for all losses without tweaking. In the ablation studies, we also test a CySGAN model without the semi-supervised segmentation loss to demonstrate its effectiveness to the framework.

\subsection{2) Augmentation-Aware Cycle Consistency:}
The U3D-BCD [5] model uses multiple training augmentations like random missing, blurry and noisy regions (Fig. 4a). We keep them in CySGAN for better segmentation quality. However, the image discriminator can easily distinguish synthesized images from real ones if the augmentations are clearly noticeable in the translated ones, breaking the balance in GAN
training. Therefore, we propose an upgraded cycle consistency (Eqn. 3) by streaming the training images for $X$ and $Y$ in both augmented and clean (unaugmented) forms. As shown in Fig. 4 (each subfigure shows consecutive slices of a 3D volume), $G$ transfers augmented $y_i$ to $\hat{x}_i$, and $F$ reconstructs $\hat{x}_i$ to $\hat{y}_i$. Instead of calculating $L_{cyc}(F,G)$ of $\hat{y}_i$ to $y_i$, we enforce its similarity to the clean $y^c_i$ (Fig. 4d). By using the augmentation-aware cycle consistency strategy, both generators learn to restore corrupted regions using 3D context$^5$ in addition to image translation. We show in the ablation studies that this strategy has a significant impact on the domain-adaptive segmentation performance.

3) Network Details and Optimization: We use 3D U-Nets [7] for $F$ and $G$. They have identical architectures, but the parameters are not shared, which is similar to CycleGAN. Each network has one input channel and four output channels for the translated image and BCD segmentation representations (Fig. 2). For the GAN objectives, we use 3D convolutional discriminators, where the image discriminators $D_{I_X}^F$ and $D_{I_Y}^G$ have a single input channel for the gray-scale images, while the segmentation-based discriminators $D_{S_X}^F$ has three input channels for the BCD representations. Each discriminator has five layers, where each one consists of a strided convolution, a batch normalization, and a non-linear activation. Following PatchGAN [16], the final layer outputs a single-channel feature map representing the realness of corresponding input patches. The idea is to evaluate the generator’s performance at the level of local image patches rather than applying a coarse global penalty. As discussed in Sec. III-B, we optimize the LSGAN objective (Eqn. 4) instead of the BCE GAN loss (Eqn. 2) for training stability. When calculating the segmentation losses, we detach the synthesized image to avoid the segmentation objectives affecting the image translation results.

We train the CySGAN model for $10^6$ iterations using the AdamW [30] optimizer with an initial learning rate of $2 \times 10^{-3}$ (decreased with cosine annealing) and batch size of 8 using 4 NVIDIA V100 GPUs. Our implementation of the proposed CySGAN framework is based on the PyTorch Connectomics [31] open-source framework.

IV. DATASETS

As discussed in related works, existing domain-adaptive segmentation models are mainly developed for 2D segmentation and semantic segmentation. To alleviate the lack of benchmark datasets for 3D domain-adaptive instance segmentation in microscopy image analysis, we also release a fully annotated dataset with dense 3D neuronal nuclei instances (Fig. 5).

1) NucExM Dataset (Target): We curated the saturated nuclei segmentation annotation for two expansion microscopy (ExM) [8] volumes by two neuroscience experts from a day 7 post-fertilization (dpf) zebrafish brain$^6$, imaged with confocal microscopy. These volumes have an anisotropic resolution of $0.325 \times 0.325 \times 2.5 \, \mu m$ in $(x,y,z)$ order, with an approximate tissue expansion factor of 7.0. Thus the effective resolution becomes $0.046 \times 0.046 \times 0.357 \, \mu m$. The two volumes are of size $2048 \times 2048 \times 255$ voxels with 9.6K and 8.8K nuclei, respectively (Table I). We downsample the volumes by $\times 4$ along $x$ and $y$ axes to $512 \times 512 \times 255$ to save computational cost during training and inference.

$^5$The strong missing-region augmentation is not applied to successive sections to facilitate using 3D context in translation and segmentation.

$^6$All procedures involving animals at the Massachusetts Institute of Technology (MIT) were conducted in accordance with the US National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the MIT Committee on Animal Care. The IACUC protocol number is 1221-100-24, which was approved on 12/23/2021.
2) Source Dataset: We use the NucMM-Z electron microscopy (EM) volume from the NucMM dataset [5] as the source data ($I_X$ and $S_X$ in Fig. 1a). The original NucMM-Z covers nearly a whole zebrafish brain at a resolution of $0.48 \times 0.48 \times 0.48 \ \mu m$. Considering the different resolutions of the source and target datasets, we crop a $200 \times 200 \times 255$ subvolume from NucMM-Z and upsample it to $512 \times 512 \times 255$ to (roughly) match the resolution. The processed volume contains 12K neuronal nuclei instances. We also apply Gaussian filtering and thresholding of the instance masks after nearest-neighbor upsampling to smooth the boundaries.

3) Datasets Comparison: Fig. 6 shows the comparison between the source (EM) and target (ExM) datasets. After downsampling of the target dataset and upsampling of the source dataset, the instance size (Fig. 6a) and nearest-neighbor distance between nuclei centers (Fig. 6b) roughly match, which is expected to help the model learn to segment 3D neuronal nuclei instances in a domain-adaptive setting. The domain gap is mainly characterized by the different intensity and contrast of object and non-object voxels (Fig. 6c).

We show in experiments that the difference in appearance can hardly be solved by traditional appearance-level adaptation approaches like histogram matching.

4) Evaluation Metric: Following common practice in instance segmentation [32], [33], we choose average precision (AP) as the evaluation metric. Specifically, for our 3D volumetric data, we choose AP-50 (i.e., AP with an IoU threshold of 0.5) and use the existing public implementation with improved efficiency for 3D volumes [21].

V. EXPERIMENTS

A. Methods in Comparison

We compare CySGAN with three types of models targeting the segmentation of a new domain without any in-domain annotation, including generalist models, appearance-level adaptation models, and feature-level adaptation models.

1) Generalist models: We compare with Cellpose [3] and StarDist [4] models using their official implementation. Cellpose predicts the flow-field representations for instances using neural networks, while StarDist predicts 3D star-convex polyhedra representations. Those models are pretrained on various training datasets covering different imaging modalities and species (e.g., the Cellpose model was pretrained on datasets with over 70k segmented objects). To improve the fairness in performance comparison, we conducted hyper-parameter tuning of the algorithms (e.g., the estimated diameters of the objects) to ensure the quality of the predictions.

2) Appearance-level adaptation: Appearance-level adaptation approaches are the models that first translate images to the target appearance for training a segmentation model. Since existing approaches are mainly developed for 2D semantic segmentation [10], [11], [13] but rarely explore 3D instance segmentation, we implemented two kinds of baseline models that conduct translation and 3D instance segmentation sequentially. Specifically, we test both histogram matching (a traditional method) and CycleGAN [9] (a deep learning-based method) as the translation module. We use U3D-BCD [5] for segmentation, which is consistent with the CySGAN generators but without the output channel for translated images. Moreover, we test the $I_X \rightarrow I_Y$ version that transfers $I_X$ to $I_Y$, and trains a model in the target domain using synthesized images, and $I_Y \rightarrow I_X$ that transfers $I_Y$ to $I_X$, and predicts the segmentation using a model trained in the source. Note

### Table I

**NUCExM DATASET METADATA.** We curated and densely annotated a neuronal nuclei segmentation dataset with two ExM volumes of zebrafish. The tissue was expanded by about $7 \times$ to increase resolution.

<table>
<thead>
<tr>
<th>Sample</th>
<th>#Volumes</th>
<th>Volume Size (each)</th>
<th>Resolution ($\mu m$)</th>
<th>Ex. Ratio</th>
<th>#Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zebrafish Brain</td>
<td>2</td>
<td>$2048 \times 2048 \times 255$</td>
<td>$0.325 \times 0.325 \times 2.5$</td>
<td>7.0</td>
<td>9.6K+8.8K</td>
</tr>
</tbody>
</table>
that $I_X \rightarrow I_Y$ adaptation is usually preferred as the $I_Y \rightarrow I_X$ approach needs to run the image translation module as inference time, introducing additional computational cost.

3) Feature-level adaptation: Appearance-level adaptation models described before first translates images between the source and target domains. In comparison, feature-level domain adaptation models commonly map the source and target distributions in the model embedding space. For feature-level domain adaptation, we implemented a model sharing a similar high-level idea as Tsai et al. [27]. Specifically, based on the same U3D-BCD model in the appearance-level adaptation models and our CySGAN, we apply the first GAN loss to match the distribution of source and target predictions (i.e., the BCD segmentation representations) and the second GAN loss to align the target features to the source features in the embedding space of the 3D U-Net model. Other training details, including data augmentations, are the same as the segmentation modules in the appearance-level adaptation models.

### B. Results

Since there are two volumes in the NucExM dataset, we only use one volume ($V_1$) to optimize the model while running inference on $V_1$ and $V_2$. The inference results of $V_2$, therefore, demonstrate the model’s generalization ability. Note that since the setting is unsupervised domain-adaptation, only the ExM images of $V_1$ are used in training without any annotations. Table II summarizes the results. Our CySGAN outperforms pretrained generalist models, feature-level adaptation models, and appearance-level adaptation models with either histogram matching or CycleGAN for image translation. Specifically, CySGAN outperforms the second-best model (CycleGAN+Segm, $I_X \rightarrow I_Y$) by absolutely 5.7%, demonstrating the effectiveness of our proposed framework. The results also show that $I_X \rightarrow I_Y$ versions generally perform better than $I_Y \rightarrow I_X$ ones in sequential models. Please note that, although the models are not optimized on $V_2$, all methods generally perform better on $V_2$ as the volume is relatively easier to segment.

The visual results in Fig. 7 show that Cellpose’s segmentation has obvious false negatives, as highlighted by the red arrows. From our hyperparameter search for Cellpose, we found that the challenging contrast of the ExM data causes missing foreground predictions. StarDist’s masks, on the other hand, tend not to align well with instance boundaries and overlap with each other, which are also highlighted using red arrows. We empirically find that the strong star-convex shape prior often overlooks other features like boundaries and thus
struggles with non-spherical shapes. Our CySGAN model that combines three predicted mask representations (Fig. 7, f-h) yields favorable 3D instance segmentation results.

C. Ablation Studies

We further validate three important design choices of CySGAN, including the data augmentations (Fig. 4), semi-supervised segmentation losses for the unlabeled domain (Eq. 7), and learning the BCD [5] representation.

Table III shows the results when removing those components from the CySGAN model on the \( V_1 \) NucExM image volume. First, without data augmentations and the corresponding cycle-consistency loss to restore corrupted regions, the performance is significantly degraded by 16.6%. We also observe that the model is prone to model collapse (i.e., the generator tends to generate a single pattern during the optimization) without data augmentations. Therefore our training strategy can improve both the performance and robustness of the domain-adaptive segmentation model. Second, CySGAN without the semi-supervised segmentation losses (which can be regarded as a 3D instance segmentation version of SUSAN [14]), the performance is decreased by 4.9% and similar to the result of the model sequentially conducting image translation and segmentation (CycleGAN + Segm in Table II). Third, we also test a model that only learns the binary foreground mask and contour map (BC), as in Wei et al. [21], without the signed distance map in the BCD representation [5]. The discriminator for the segmentation-based GAN loss is updated accordingly to have two input channels without modifying other training protocols. The BC version is worse than the default CySGAN model by 8.4%, validating the importance of the signed distance map in segmenting closely-touching 3D instances. Those results demonstrate the essentiality of those components in CySGAN and also provide informative data points to quantify the importance of those designs.

VI. Conclusion

In this work, we present CySGAN, a unified domain-adaptive segmentation framework optimized with image translation losses as well as supervised and semi-supervised instance segmentation losses to tackle an unlabeled imaging modality. CySGAN outperforms and simplifies models that conduct translation and segmentation using separate networks. We also publicly release the NucExM dataset as a testbed for future domain-adaptive 3D instance segmentation models. In our application scenario, the morphology of the source and target objects are relatively close. Thus, important future directions include segmenting modalities where the instance structures differ significantly from those in the source domain.

### Table III

<table>
<thead>
<tr>
<th>Configuration</th>
<th>w/o Augmentation</th>
<th>w/o Semi-sup Losses</th>
<th>w/ BC only</th>
<th>CySGAN (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP-50 (( V_1 ))</td>
<td>0.761 (-0.166)</td>
<td>0.878 (-0.049)</td>
<td>0.843 (-0.084)</td>
<td>0.927</td>
</tr>
</tbody>
</table>

**References**


