Vox-UDA: Voxel-wise Unsupervised Domain Adaptation for Cryo-Electron Subtomogram Segmentation with Denoised Pseudo-Labeling

Haoran Li[1, 2], **Xingjian Li**[3], **Jiahua Shi**[4], **Huaming Chen**[5], **Bo Du**[6], **Daisuke Kihara**[7], **Johan Barthelemy**[8], **Jun Shen**[1*], **Min Xu**[3*]

[1]School of Computing and Information Technology, University of Wollongong, Australia
[2]ARC Training Centre for Innovative Composites for the Future of Sustainable Mining, University of Wollongong, Australia
[3]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University, USA
[4]Centre for Nutrition and Food Sciences, University of Queensland, Australia
[5]School of Electrical and Computer Engineering, University of Sydney, Australia
[6]Department of Business Strategy and Innovation, Griffith University, Australia
[7]Department of Biological Sciences, Purdue University, USA
[8]NVIDIA, USA
hl644@uowmail.edu.au, lixj04@gmail.com, jiahua.shi@uq.edu.au, huaming.chen@sydney.edu.au, bo.du@griffith.edu.au, dkihara@purdue.edu, jbarthelemy@nvidia.com, jshen@uow.edu.au, mxu1@cs.cmu.edu

Abstract

Cryo-Electron Tomography (cryo-ET) is a 3D imaging technology that facilitates the study of macromolecular structures at near-atomic resolution. Recent volumetric segmentation approaches on cryo-ET images have drawn widespread interest in the biological sector. However, existing methods heavily rely on manually labeled data, which requires highly professional skills, thereby hindering the adoption of fully-supervised approaches for cryo-ET images. Some unsupervised domain adaptation (UDA) approaches have been designed to enhance the segmentation network performance using unlabeled data. However, applying these methods directly to cryo-ET image segmentation tasks remains challenging due to two main issues: 1) the source dataset, usually obtained through simulation, contains a fixed level of noise, while the target dataset, directly collected from raw-data from the real-world scenario, have unpredictable noise levels. 2) the source data used for training typically consists of known macromoleculars. In contrast, the target domain data are often unknown, causing the model to be biased towards those known macromolecules, leading to a domain shift problem. To address such challenges, in this work, we introduce a voxel-wise unsupervised domain adaptation approach, termed Vox-UDA, specifically for cryo-ET subtomogram segmentation. Vox-UDA incorporates a noise generation module to simulate target-like noises in the source dataset for cross-noise level adaptation. Additionally, we propose a denoised pseudo-labeling strategy based on the improved Bilateral Filter to alleviate the domain shift problem. More importantly, we construct the first UDA cryo-ET subtomogram segmentation benchmark on three experimental datasets. Extensive experimental results on multiple benchmarks and newly curated real-world datasets demonstrate the superiority of our proposed approach compared to state-of-the-art UDA methods.

Code — https://github.com/xulabs/aitom.

Introduction

Cryo-Electron Tomography (cryo-ET) is one cutting-edge imaging technique which enables three-dimensional views of biological samples in a native frozen-hydrated state (Oikonomou and Jensen 2017). This automatic electron tomography technique allows biologists to capture high-resolution structures of macromolecular complexes (Wan and Briggs 2016), which plays an important role in the field of drug discovery and disease treatment. Inspired by the development of deep learning research in recent years, some efforts have been made in cryo-ET image analysis, especially for the subtomogram segmentation task (Zhu et al. 2021; Zhou et al. 2021). Subtomogram segmentation is a 3D segmentation task that aims to mine the meaningful information of the target macromolecular on the voxel-level. However, existing methods heavily rely on manual annotations which are highly subjective and resource-intensive.

To tackle the challenges for data annotation, the classical unsupervised domain adaptation (UDA) methods involve transferring the knowledge from labeled source domains to unlabelled target domains (Liu et al. 2020; Zhao et al. 2022; Zhang et al. 2023a). Most existing approaches are designed for 2D images, making them less effective for 3D tasks. While some recent works explore UDA on 3D images (Shin et al. 2023; Xian et al. 2023), they typically convert 3D input into 2D slices for network processing, resulting in a loss of spatial information. MAPSeg (Zhang et al. 2024) proposes a voxel-level pseudo-labeling method based on MAE. However, label distortion remains a main challenge for UDA.

In this paper, we introduce a novel UDA approach using large simulated macromolecular data (Eisenstein, Danev, and Pilhofer 2019; Hagen, Wan, and Briggs 2017) as the source domain dataset and experimental dataset as the target domain dataset. Advances in data simulation techniques have made it possible to generate cryo-ET subtomogram data beyond traditional biological methods. Given the structure of macromolecules, existing generative methods (Martinez-Sanchez et al. 2024; Harar et al. 2023) can

[*]Corresponding Authors.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

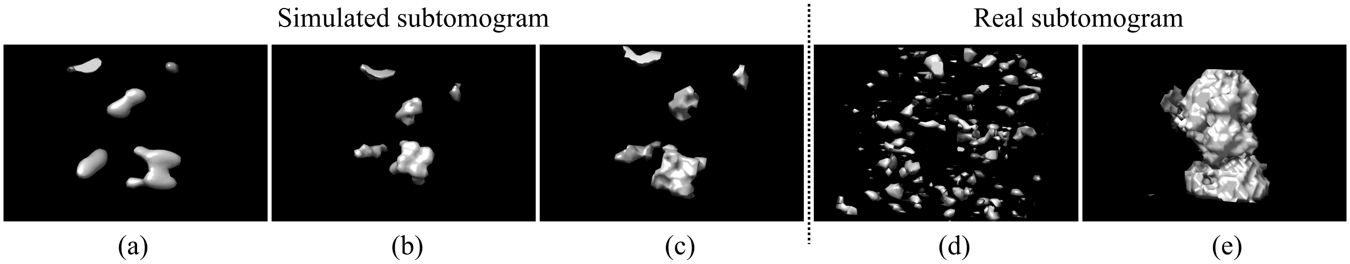

Figure 1: Some examples of the subtomograms and their corresponding segmentation masks. This figure shows: (a) simulated 3D cryo-ET subtomogram; (b) grey-scale ground-truth segmentation mask; (c) binary segmentation mask after pre-processd (b), we set a threshold (300 in this paper) to turn the grey-scale mask into a binary one; (d) and (e) are experimental 3D cryo-ET subtomogram and its binary mask, respectively.

create realistic synthetic datasets with corresponding voxel-level segmentation masks, offering a cost-effective alternative to traditional methods that require high-end equipment and significant human expertise. Nevertheless, the substantial differences between two domains introduce new challenges for the UDA task. Firstly, the simulated data is generated with fixed parameters, yielding a consistent noise level in each subtomogram (typically 0.03 dB or 0.05 dB), whereas the noise rate is unpredictable in the experimental dataset. Some examples of subtomograms are shown in Figure. 1. Secondly, although subtomogram segmentation is a binary task, the molecular categories in simulated subtomograms often differ from those in experimental ones, which will cause the segmentation model to be biased towards the simulated ones and lead to the domain shift problem.

To address the aforementioned challenges, we propose a voxel-wise UDA framework, termed Vox-UDA, for cryo-ET subtomogram segmentation. Vox-UDA comprises two key components: a noise generation module (NGM) and a denoised pseudo-labeling (DPL) strategy. NGM generates Gaussian noise from a subset of the target dataset and applies it to the source samples, creating a target-like noisy effect. Meanwhile, DPL proposes an enhanced bilateral filter that replaces pixel differences with gradient difference, making it more effective for 3D grayscale images. DPL preserves edge information during denoising as much as possible, resulting in undistorted pseudo-labels. These pseudo-labels provide additional supervision signals to mitigate the domain shift problem, thereby enhancing the model's performance on the target data.

In summary, our contributions are as follows: 1) To the best of our knowledge, we are the first to establish a paradigm for voxel-wise UDA segmentation in cryo-ET images (termed Vox-UDA). 2) Vox-UDA incorporates a noise generation module (NGM) and a denoised pseudo-labeling (DPL) strategy to enable the simulation of a target-like noisy effect, and it provides additional supervision signals to address domain shift. 3) We further propose an improved bilateral filter that, by being sensitive to the gradient changes, preserves edge information as much as possible while eliminating noises to obtain high-quality pseudo-labels. 4) We provide the first benchmark for UDA in cryo-ET subtomogram segmentation across three experimental datasets, including two newly curated datasets that are publicly available for the first time.

Related Work

Unsupervised Domain Adaptation for Vision Tasks Under the unsupervised domain adaptation (UDA) settings, there are two types of datasets being used for training: the source domain dataset which is fully labelled, and the target domain dataset, which is unlabelled. The first UDA approach is proposed by (Ganin et al. 2016), which aims to transfer the model trained on source data to target data without introducing additional annotations through adversarial learning. Since UDA greatly expands the model's generalization ability, the model can be adapted to new domains without requiring labeled data in the target domain and is introduced into various tasks. As cryo-ET subtomogram segmentation is a segmentation task, in this paper, we mainly focus on the UDA approaches applied in semantic segmentation tasks. For segmentation, the UDA methods aim to eliminate the cross-domain discrepancies through the content at both the feature- (Zou et al. 2018) and pixel-level (Zheng et al. 2023) and, have achieved excellent results in medical image segmentation (Zhang et al. 2023b; Xie et al. 2022; Li et al. 2022). However, although these methods achieve great performance in UDA segmentation, they are primarily designed for 2D images, which are not suitable for cryo-ET subtomogram segmentation as tomographies are often volumetric images. To handle the UDA challenge in 3D segmentation tasks, Shin et al. (Shin et al. 2023) proposed a cross-modality translation method to generate synthetic 3D target volumes from source 2D scans. Xu et al. (Xu et al. 2023) applied a fast Fourier transform to convert input 2D slices into the frequency domain. A consistency loss was utilized to simultaneously constrain both the feature domain and frequency domain to achieve UDA. In this work, we focus on the noise differences between the simulated subtomogram and experimental subtomogram data, aiming to enhance the model's robustness to the noises.

Cryo-Electron Tomography Cryo-electron tomography (cryo-ET) integrates cryogenic specimen preparation, electron microscopy for data acquisition, and tomographic reconstruction for 3D visualization (Koning 2010). A cryo-

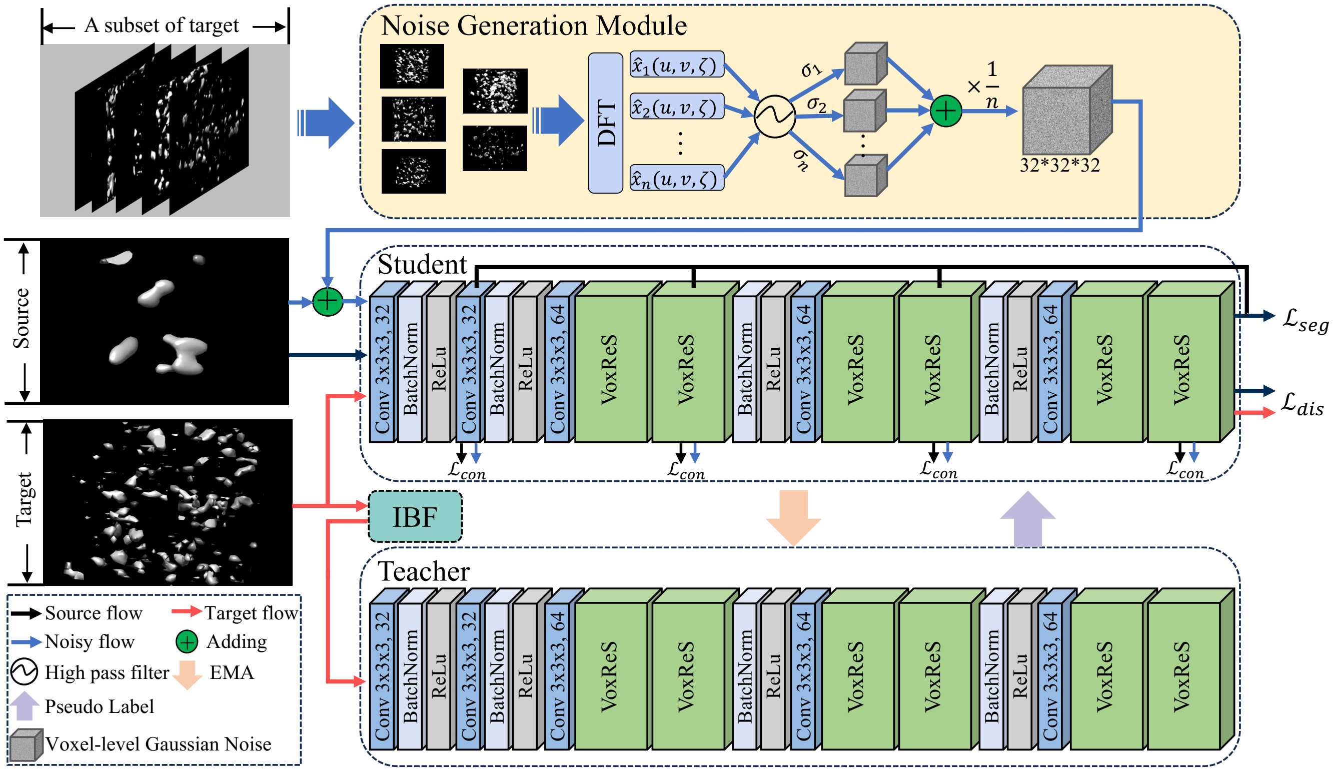

Figure 2: Overview of our proposed Vox-UDA framework. IBF denotes the improved Bilateral Filter. We use different colors to represent different flows. Best viewed in color.

ET subtomogram is a small cubic sub-volume extracted from a tomogram, normally only a single macromolecular complex is contained in each subtomogram. Inspired by recent advancements in deep learning, their applications on cryo-ET have drawn widespread interest with their potential to aid in the corresponding cryo-ET tasks, *e.g.*, subtomogram alignment (Zeng and Xu 2020), subtomogram classification (Wan, Khavnekar, and Wagner 2024; Bandyopadhyay et al. 2022) and subtomogram segmentation (Zhu et al. 2021). However, deep learning methods rely on large amounts of data annotation, which is particularly challenging for cryo-ET images. In this paper, we will propose an UDA approach for subtomogram segmentation, which aims at utilizing a large amount of cost-free annotated simulated data for knowledge transfer, enabling the segmentation network to generalize on experimental cryo-ET subtomograms.

Method

Our proposed framework is based on VoxResNet (Chen et al. 2018), a state-of-the-art method designed for fully-supervised voxel-level segmentation. Given a source domain dataset $\mathcal{S} = \{x_i^s, y_i^s\}_{i=1}^N$ and a target domain dataset $\mathcal{T} = \{x_j^t\}_{j=1}^M$, where x_i represents the input 3D subtomogram and y_i denotes the 3D ground-truth mask, we aim to train a voxel segmentation network for the target domain only using ground-truth supervision signals from the source domain. Figure. 2 illustrates the details of the proposed Vox-UDA. As can be seen from the figure, Vox-UDA takes x_i^s, x_i^t and a subset of $\mathcal{T}$ as input. This subset $\mathcal{T}_{N_{sampled}}$ is randomly sampled from $\mathcal{T}$, which contains $N_{sampled}$ samples. $\mathcal{T}_{N_{sampled}}$ is then sent to the noise generation module (NGM) to obtain the target-like voxel-wise Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_t^2\mathbf{I})$. Further, ϵ is introduced to the source input x_i^s to produce updated input $x_i^{s'}$. x_i^s, $x_i^{s'}$ and x_j^t are all passed to the student network to acquire segmentation loss $\mathcal{L}_{seg}$, consistency loss $\mathcal{L}_{con}$ and discriminator loss $\mathcal{L}_{dis}$ for optimization. Hence, the overall loss can be rewritten as

$$\mathcal{L} = \mathcal{L}_{seg} + \mathcal{L}_{con} + \mathcal{L}_{dis}. \tag{1}$$

Furthermore, to handle the domain shift problem, we design a denoised pseudo-labeling strategy. x_j^t is sent to the improved Bilateral Filter (IBF) to eliminate its noise and then sent to the teacher network to obtain the pseudo-label, which is then used to tune the student network for better performance. Note that the threshold η used for pseudo-label selection is set to 0.85 and the teacher network is updated via exponential moving average (EMA).

Noise Generation Module

Given a sample x_n^t from the input $\mathcal{T}_{N_{sampled}}$, we first apply Discrete Fourier Transform (DFT) to obtain its frequency information

$$\hat{x}_n(u, v, \zeta) = \xi[x^t], \tag{2}$$

where u, v and ζ represent the spatial frequencies of the Fourier transform, and ξ denotes the Discrete Fourier transform. In the frequency domain, low-frequency information corresponds to the textural details of the target object, while large amounts of noise with little edge information about the object are usually encompassed in the high-frequency information. To obtain the noise encompassed in the high-frequency information, $x_n(u, v, \zeta)$ is then passed to a high-pass filter to eliminate the textural details contained in low-frequency information

$$\hat{x}'_n(u, v, \zeta) = H_{high}(u, v, \zeta)\hat{x}_n(u, v, \zeta), \quad (3)$$

where $H_{high}(u, v, \zeta)$ denotes the high-pass filter. The filter rate is set to 24.4%, which means only 24.4% remains while the rest of them are filtered. Inverse Discrete Fourier transform (iDFT) is further applied to recover voxel-level information from the filtered frequency domain $x'_n(u, v, \zeta)$:

$$x_n^{t'} = \xi^{-1}[\hat{x}'_n], \quad (4)$$

where $\xi^{-1}[\cdot]$ denotes the Inverse Discrete Fourier Transform. As discussed in the first section, the noise level of each input from the target domain is unpredictable, hence instead of using the noise from single x_n^t, we calculate the average noise level $\overline{x}_n^{t'}$ from the whole subset. Since deep learning models are more sensitive to noise that conforms to a probability distribution (Lehtinen et al. 2018), we set the Gaussian noise as the input noise for noise generation (See detailed discussions in Appendix B.4). Therefore, instead of directly introducing $\overline{x}_n^{t'}$ to the source input x_i^s, we only take its variance σ_t^2 and generate a Gaussian noise based on

$$\epsilon = \mathcal{N}(0, \sigma_t^2\mathbf{I}), \quad (5)$$

where $\mathcal{N}(0, \sigma_t^2\mathbf{I})$ denotes a random generated Gaussian noise with expectation equals to 0, and variance equals to σ_t^2. And the updated source input is obtained through

$$x_i^{s'} = x_i^s + \epsilon. \quad (6)$$

x_i^s and $x_i^{s'}$ are both sent to the student network to obtain the consistency loss $\mathcal{L}_{con}$. Following VoxResNet, we also take the output feature embeddings from the same layers as the input of the loss function

$$\begin{aligned}\mathcal{L}_{con} = \lambda_1\mathcal{L}_{BN}(f_c, f'_c) + \lambda_2\mathcal{L}_{BN}(f_{v2}, f'_{v2}) \\ +\lambda_3\mathcal{L}_{BN}(f_{v4}, f'_{v4}) + \lambda_4\mathcal{L}_{BN}(f_{v6}, f'_{v6}),\end{aligned} \quad (7)$$

where $\mathcal{L}_{BN}$ denotes the cosine similarity loss and λ_n denotes the weights to control the relative importance among different consistency losses. Although NGM is introduced to simulate target-like noises, it is impossible to create a noise environment in the source domain that is entirely the same as the target domain. On the other hand, while the shallower layers of the decoder contain more textural information, the deeper layers contain more edge information (Chen et al. 2017). To overcome these constraints, our solution is, instead of an equal superposition, we assign different weights to the different layers to control the weighting of texture and edge consistency losses.

Denoised Pseudo-Labeling

Although the NGM can narrow noise level gaps between two domains, as aforementioned, the segmentation network is still biased toward the source data due to the domain shift problem. Therefore, we provide an extra supervision signal for optimization through pseudo-labeling. However, due to the noise level being unknown in the target domain and also that such noise may lead to distorted pseudo-labels further harming the performance of the model, we propose a denoised pseudo-labeling strategy instead. Unlike the existing pseudo-labeling method whereby adding an extra training step, we use the student-teacher structure (Sohn et al. 2020). Before x_j^t is sent to the teacher network to obtain the pseudo-label, we first perform denoising on x_j^t. We designed three different denoising methods: $1)$ directly using NGM for denoising, $2)$ using Bilateral Filter for denoising, and $3)$ using our designed improved Bilateral Filter (IBF) for noise reduction.

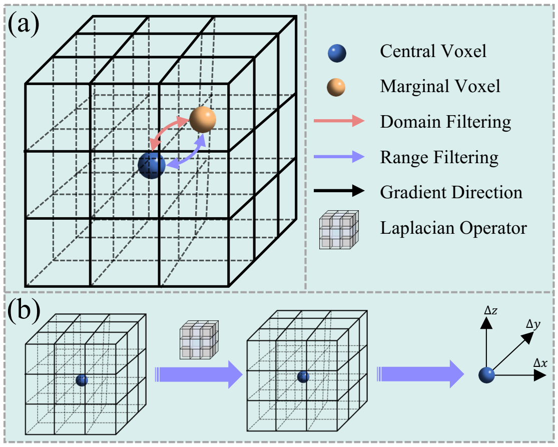

Figure 3: Proposed improved Bilateral Filter. (a) Both domain filtering and range filtering are applied to a sub-figure extracted from the input target subtomogram with size 3 × 3 × 3. (b) Deploying Laplace transform to obtain the gradient changes used in range filtering.

NGM Denoising. The x_j^t is directly sent to the NGM to obtain its noise $x_j^{t'}$. Hence, the denoised image can be represented as $\widetilde{x}_j^t = \left(x_j^t - x_j^{t'}\right)$.

Bilateral Filter Denoising. Although noise can be partially removed through frequency domain analysis, some edge information will also be eliminated, leading to distortion of the pseudo-labels. Therefore, we further deploy a non-linear approach, Bilateral Filter (Elad 2002), as the denoiser instead of the NGM. Bilateral filter (BF) consists of a domain Gaussian kernel and a range Gaussian kernel, the former is used to eliminate the noises, and the latter is to retain edge information as much as possible during filtering. BF uses a sliding window, extracting a $3 \times 3 \times 3$ sub-figure for filtering operations each time. Given the central voxel v_p and the rest of voxels $v_q, q \in V$ of the sub-figure, the updated voxel can

	Source macromoleculars: 1bxn, 1f1b, and 1yg6							
Method	mIoU	mIoU$_{ribo}$	mIoU$_{26S}$	mIoU$_{TRiC}$	Dice	Dice$_{ribo}$	Dice$_{26S}$	Dice$_{TRiC}$
w/o adaptation	9.7	11.1	2.6	2.6	17.4	19.9	5.1	5.0
Fully Supervised	55.8	57.9	36.7	48.8	71.0	73.1	53.0	63.1
DANN (Ganin et al. 2016)	38.4	43.0	6.5	11.7	53.0	59.1	11.2	16.8
PDAM (Liu et al. 2020)	39.8	43.3	13.8	22.6	55.1	59.6	21.5	31.9
ASC (Xu et al. 2023)	40.4	43.4	19.2	23.3	55.8	59.7	28.7	32.7
LE-UDA (Zhao et al. 2022)	41.5	44.7	18.4	23.9	56.8	61.0	28.4	32.6
MAPSeg (Zhao et al. 2024)	44.6	46.4	19.7	25.1	60.0	64.1	30.1	33.9
Vox-UDA(w NGM)	48.5	50.6	**32.2**	38.6	64.4	66.8	**47.1**	51.5
Vox-UDA(w BF)	49.1	50.4	30.5	39.2	64.5	67.1	46.9	50.7
Vox-UDA(w IBF)	**50.3**	**53.8**	28.8	**41.3**	**65.9**	**68.5**	44.0	**52.8**

Table 1: Comparison of experimental results on UDA cryo-ET subtomogram segmentation between Vox-UDA and baselines using **Poly-GA** as target dataset. Best results are in bold font.

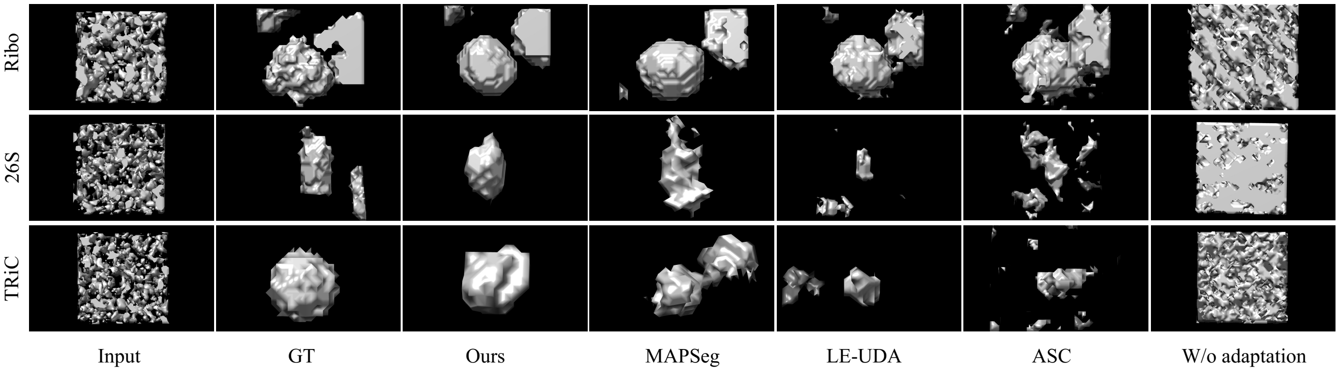

Figure 4: Visualization of subtomogram segmentation results using 1bxn, 1f1b, and 1yg6 as the source datasets. We use UCSF Chimera (Pettersen et al. 2004) for 3D cryo-ET visualization.

be represented as

$$v_q^{'} = \text{BF}(v_q) = \frac{\sum_{q \in V} G_{\sigma_d}(||p-q||) G_{\sigma_r}(v_q - v_p) \times v_q}{\sum_{q \in V} G_{\sigma_d}(||p-q||) G_{\sigma_r}(v_q - v_p)}, \tag{8}$$

where $||\cdot||$ denotes the Euclidean distance (Wang, Zhang, and Feng 2005), σ_d and σ_r denote the domain hyperparameter and range hyperparameter, and G_σ denotes the Gaussian kernel

$$G_\sigma(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{2\pi\sigma}. \tag{9}$$

Hence, the denoised image can be represented as $\widetilde{x}_j^t = \text{BF}(x_j^t)$.

IBF Denoising. The key point of the Bilateral Filter is the design of using two separate Gaussian kernels for different tasks. However, the range kernel introduced for retaining edge information mainly focuses on the voxel-level color difference, which indeed can achieve satisfactory results in the RGB space, but in the grayscale space, there might be more differences in brightness, which could affect the effectiveness of this kernel. Therefore, we further propose an improved Bilateral Filter (IBF), which uses the gradient of each voxel instead of its value for the range kernel for edge retaining. In detail, we reflect Laplace operator (Van Vliet, Young, and Beckers 1989) into 3-dimension and calculate the gradient of each voxel v_q in the h, w, and d (height, width, and depth) directions in three-dimensional space. Since voxel space is discrete, the gradient of v_q in each direction can be represented as

$$\frac{\partial \Delta}{\partial \vec{v}_q^h} = \Delta(v_{q+1}^h, v_q^w, v_q^d) - \Delta(v_{q-1}^h, v_q^w, v_q^d), \tag{10}$$

$$\frac{\partial \Delta}{\partial \vec{v}_q^w} = \Delta(v_q^h, v_{q+1}^w, v_q^d) - \Delta(v_q^h, v_{q-1}^w, v_q^d), \tag{11}$$

$$\frac{\partial \Delta}{\partial \vec{v}_q^d} = \Delta(v_q^h, v_q^w, v_{q+1}^d) - \Delta(v_q^h, v_q^w, v_{q-1}^d), \tag{12}$$

where x_q^h, x_q^w and x_q^d denote the values of v_q in the h, w, and d directions, and Δ denotes the Laplace operator. Moreover, compared to the gradient of the central voxel v_p, if a voxel is belonging to the object (inside), their gradient should be similar. Otherwise, the difference between the two gradient should be large. Therefore, we can replace the second filter in Eq 8 and obtain the improved Bilateral Filter (IBF):

$$v_q^{'} = \text{IBF}(v_q) = \frac{\sum_{q \in V} G_{\sigma_d}(||p-q||) G_{\sigma_r}(\frac{\partial \Delta}{\partial \vec{v}_p} - \frac{\partial \Delta}{\partial \vec{v}_q}) \times v_q}{\sum_{q \in V} G_{\sigma_d}(||p-q||) G_{\sigma_r}(\frac{\partial \Delta}{\partial \vec{v}_p} - \frac{\partial \Delta}{\partial \vec{v}_q})}, \tag{13}$$

where

$$\frac{\partial \Delta}{\partial \vec{v}_p} - \frac{\partial \Delta}{\partial \vec{v}_q} = (\frac{\partial \Delta}{\partial \vec{v}_p^h}, \frac{\partial \Delta}{\partial \vec{v}_p^w}, \frac{\partial \Delta}{\partial \vec{v}_p^d}) - (\frac{\partial \Delta}{\partial \vec{v}_q^h}, \frac{\partial \Delta}{\partial \vec{v}_q^w}, \frac{\partial \Delta}{\partial \vec{v}_q^d}). \tag{14}$$

	Source macromoleculars: 2byu, 2h12, and 21db							
Method	mIoU	mIoU$_{ribo}$	mIoU$_{26S}$	mIoU$_{TRiC}$	Dice	Dice$_{ribo}$	Dice$_{26S}$	Dice$_{TRiC}$
w/o adaptation	12.7	14.2	3.7	3.0	22.2	24.7	7.2	5.8
Fully Supervised	55.8	57.9	36.7	48.8	71.0	73.1	53.0	63.1
DANN (Ganin et al. 2016)	31.9	36.1	3.8	6.2	45.9	51.9	6.4	8.7
PDAM (Liu et al. 2020)	39.1	43.1	10.9	15.7	54.1	59.4	17.7	24.0
ASC (Xu et al. 2023)	41.7	45.2	24.7	13.3	56.9	61.2	38.1	19.8
LE-UDA (Zhao et al. 2022)	43.1	46.4	21.9	22.3	58.6	62.6	33.8	32.4
MAPSeg (Zhao et al. 2024)	44.2	47.7	22.6	23.2	59.9	63.9	34.5	33.4
Vox-UDA(w NGM)	47.5	50.1	27.5	34.7	63.2	66.3	41.0	46.8
Vox-UDA(w BF)	48.0	50.3	27.8	34.4	63.8	66.7	39.9	47.0
Vox-UDA(w IBF)	**49.5**	**52.4**	**28.3**	**35.1**	**65.2**	**68.9**	**41.3**	**47.7**

Table 2: Experimental results on UDA cryo-ET subtomogram segmentation under different dataset settings. Different from the results reported in Table 1, we set 2byu, 2h12, and 21db as the source datasets and using Poly-GA as target dataset.

	Setting I		Setting II	
Method	mIoU	Dice	mIoU	Dice
w/o adaptation	37.8	54.8	38.1	55.2
Fully Supervised	99.1	99.5	99.1	99.5
DANN (Ganin et al. 2016)	45.4	62.5	62.3	76.8
PDAM (Liu et al. 2020)	82.0	90.1	78.2	87.8
ASC (Xu et al. 2023)	87.9	93.5	91.3	95.4
LE-UDA (Zhao et al. 2022)	92.9	96.3	93.0	96.4
MAPSeg (Zhao et al. 2024)	94.9	97.4	94.0	96.9
Vox-UDA(w NGM)	94.4	97.1	94.1	97.0
Vox-UDA(w BF)	95.0	97.5	95.1	97.5
Vox-UDA(w IBF)	**98.9**	**99.4**	**97.4**	**98.7**

Table 3: Experimental results on UDA cryo-ET subtomogram segmentation using ***Mycoplasma pneumoniae*** as target dataset.

Consequently, the denoised image is denoted as $\widetilde{x}_j^t = \text{IBF}(x_j^t)$. $\widetilde{x}_j^t$ is further sent to the teacher network and obtain the pseudo-label with the threshold η. The pseudo-label is further sent back to the student network as a supervision signal for the target flow.

Experiments

Experimental Settings

Implementation Details and Hyperparameters We utilize the VoxResNet as our base architecture. The whole model is trained on a single NVIDIA A100 Tensor Core GPU with 80GB memory. For training, we choose the Adam optimizer with an initial learning rate set to 1e-3 for optimization. The model is trained for 300 epochs with batch size of 16. The learning rate is decayed by 90% every 100 epochs. The hyperparameters sampled number $N_{sampled}$ and filter rate ρ are empirically set to 10 and 24.4%, separately. For the improved Bilateral Filter, the domain hyperparameter σ_d and range hyperparameter σ_r are set to 120 and 1.2, respectively. We use Sobel operator as the Laplacian operator.

Datasets and Evaluation Metrics We conduct experiments on two types of datasets: simulated dataset (source dataset) and experimental dataset (target dataset).

Simulated Dataset The simulated dataset used as the source dataset is generated following the same generation process as (Zeng et al. 2023). We choose six representative macromolecule complexes in our simulated datasets and divide them into two groups as two separate source datasets (1bxn, 1f1b, and 1yg6; 2byu, 2h12, and 21db). For each macromolecule complex, we simulate it with two different noise levels, with SNR of 0.03 and 0.05, and each of them contains 500 samples. Following existing work (Zeng and Xu 2020; Liao et al. 2020), all the input subtomogram are resized to 32^3. The simulated dataset contains 6,000 samples in total (3,000 samples for each source dataset).

Experimental Dataset We use three experimental datasets as the target datasets: Poly-GA (Guo et al. 2018), *Mycoplasma pneumoniae* (O'Reilly et al. 2020) and Erwinia (Prichard et al. 2023). **Poly-GA** is a public dataset, which contains 1,033 samples in total. Each subtomogram is also re-scaled to size 32^3. *Mycoplasma pneumoniae* and Erwinia are collected from the public repository EMDB (wwPDB Consortium 2024). ***Mycoplasma pneumoniae*** contains 10 samples in total. Each subtomogram is resized to 192^3 and cut into multiple non-overlapping patches with size 32^3 as the inputs for the model. **Erwinia** also contains 10 subtomogram samples, and each of them is resized to 64^3 and also cut into multiple non-overlapping patches with size 32^3. Hence, there are 2160 samples in *Mycoplasma pneumoniae* dataset and 80 samples in the Erwinia dataset (See detailed descriptions in Appendix A).

For evaluation, the mean intersection of union (**mIoU**) and dice similarity coefficient (**Dice**) are employed to evaluate the segmentation performance.

Baselines We implement the most relevant traditional and state-of-the-art UDA approaches for 2D image segmentation on our task, including single discriminator-based (DANN (Ganin et al. 2016)) and image synthesizing based (PDAM (Liu et al. 2020)). And we also include the three most recent approaches designed for volumetric images, including ASC (Xu et al. 2023), LE-UDA (Zhao et al. 2022) and MAPSeg (Zhang et al. 2024). ASC and LE-UDA cut 3D images into 2D slices and apply UDA on 2D scenario, while MAPSeg directly performs UDA in 3D voxel space.

Source macromoleculars: 1bxn, 1f1b and 1yg6														
N$_{sample}$	mIoU	Dice	ρ	mIoU	Dice	$\lambda_1 \rightarrow \lambda_4$	mIoU	Dice	σ_d	mIoU	Dice	σ_r	mIoU	Dice
5	41.2	57.1	8.4%	41.7	57.3	[0.1, 0.1, 0.4, 0.4]	44.4	60.2	100	49.3	64.9	0.8	47.5	63.2
10	**50.3**	**65.9**	17.8%	43.5	59.6	**[0.2, 0.2, 0.3, 0.3]**	**50.3**	**65.9**	**120**	**50.3**	**65.9**	1.0	48.0	63.8
15	43.6	59.4	**24.4%**	**50.3**	**65.9**	[0.3, 0.3, 0.2, 0.2]	45.3	61.0	140	49.1	64.5	**1.2**	**50.3**	**65.9**
20	42.6	58.3	42.2%	41.0	56.8	[0.4, 0.4, 0.1, 0.1]	42.2	57.8	160	47.0	62.5	1.4	49.5	65.2

Table 4: Ablation study on the hyperparameters of our proposed model on Poly-GA dataset. N$_{sample}$, ρ and λ_n denote the sampled number of target data used for noise generation, the high-pass filter rate and the weights for different losses, respectively.

We also set a "w/o adaptation" setting and a "Fully Supervised" setting for comparison. The "w/o adaptation" setting is an original VoxResNet trained on the source dataset without adaptation. The "Fully Supervised" setting is a VoXResNet trained on the target datasets with all labels provided, as the upper bound.

Comparisons with State-of-the-Arts

We report the segmentation results on the Poly-GA dataset in Table 1 using the [1bxn, 1f1b, and 1yg6] as the source dataset. As can be observed in the table, our approach outperforms all the state-of-the-art methods. Compared with "w/o adaptation", DANN and PDAM indeed boost the model's performance, however, the effect is not obvious compared with our Vox-UDA (w IBF) (*i.e.,* PDAM achieves 55.1 in $Dice$ while Vox-UDA (w IBF) achieves 65.9). And compared with three recent UDA methods, our proposed Vox-UDA (w IBF) still excels on target subtomogram segmentation, which leads to significant improvements in both $mIoU$ (*i.e.,* 41.5 → 50.3) and $Dice$ (*i.e.,* 60.0 → 65.9). We also report extra UDA setting results in Table 2 by using the other three macromoleculars [2byu, 2h12, and 21db] as source datasets, by which our proposed method still achieves state-of-the-art performance over all the comparison approaches. Figure 4 shows the visualization of the segmentation results on the Poly-GA dataset using 1bxn, 1f1b, and 1yg6 as source dataset.

We also report the segmentation results on the *Mycoplasma pneumoniae* dataset in Table 3 following the same experimental settings as Table 1 and Table 2. As can be seen, our Vox-UDA performs better over all SOTAs, *e.g.,* $mIoU$ increased by 4.2% and $Dice$ increased by 1.9%. For visualization results on *Mycoplasma pneumoniae* dataset and results on the **Erwinia** dataset, please refer to Appendix B.1.

Ablation Study

Effectiveness of the Improved Bilateral Filter To validate the effectiveness of the proposed improved Bilateral Filter (IBF) for denoised pseudo-labeling, we conduct experiments using the three different denoisers respectively, and report the segmentation results in both Table 1 and Table 2. As can be seen from the tables, comparing the method of using NGM as a denoiser, employing Bilateral Filtering (BF) indeed brings a performance improvement (*i.e.,* $mIoU$ increased 1.2% in Table 1 and $Dice$ increased 1.0% in Table 2). This is because BF can preserve some edge information while denoising, thereby avoiding pseudo-label distortion. However, as discussed in the third section, the range kernel of BF is not suitable for grayscale inputs. Our proposed IBF addresses this drawback by using a Laplacian transform, which allows the range kernel to focus more on gradient changes in the voxel space rather than value changes. Therefore, our new model achieves the best performance via the proposed IBF (*i.e.,* $mIoU$ increased 2.4% in Table 1 and $Dice$ increased 3.3% in Table 2).

Effectiveness of Different Proposed Modules We evaluate our Vox-UDA following the same experimental setting in Table 1 for the ablation study and use Vox-UDA (w IBF) as the final result of our proposed method. Experimental results demonstrates the effectiveness of our proposed two modules in dealing with the challenges for UDA in subtomogram segmentation. In the meantime, compared with using these two modules only, Vox-UDA achieves a significant performance improvement (See detailed discussions in Appendix B.5).

Hyperparameter Analysis We herein further evaluate the hyperparameters in our approach. As shown in Table 4, we evaluate the sampled number N_{sample}, the high-pass filter rate ρ, the weight λ_n for consistency losses, the domain hyperparameter σ_d and range hyperparameter σ_r. Experimental results demonstrate that our settings are the most reasonable. Whether increasing (decreasing) the number of sampled targets, expanding (reducing) the filter range, placing more emphasis on the convergence of consistency loss towards textural information, or increasing or altering σ_d and σ_r, would not lead to an improvement in the model's performance (See detailed discussions in Appendix B.6).

Conclusion

In this paper, we propose a voxel-level unsupervised domain adaptation approach, termed Vox-UDA, for the subtomogram segmentation task. In detail, our Vox-UDA consists of a Noise Generation Module (NGM) and a denoised pseudo-labeling (DPL) strategy. NGM generates target-like Gaussian noise for the source domain data, while DPL uses an improved bilateral filter (IBF) to provide denoised target domain data for pseudo-labeling to boost the segmentation performance. We have conducted large-scale experiments to demonstrate the prominent performance of our method and provide the first benchmark for UDA of cryo-ET subtomogram segmentation task. We anticipate our novel method can contribute more to the research in cryo-ET in terms of methodology and possibly enhanced interpretability. Furthermore, we would propose that future research endeavors focus on enhancing the scalability of our method for a broader range of biomedical 3D image segmentation tasks.

Acknowledgments

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629, DBI-2238093, IIS-2007595, IIS-2211597, and MCB-2205148, UPMC Enterprises, Oracle for Research, and AMD HPC Fund.

The authors acknowledge NVIDIA and its research support team for the help provided to conduct this work. This work was also partially supported by the Australian Research Council (ARC) Industrial Transformation Training Centres (ITTC) for Innovative Composites for the Future of Sustainable Mining Equipment under Grant IC220100028.

DK acknowledges fundings from NIH (R01GM133840) and from NSF (IIS2211598, MCB1925643, DMS2151678, DBI2146026, and DBI2003635).

References

Bandyopadhyay, H.; Deng, Z.; Ding, L.; Liu, S.; Uddin, M. R.; Zeng, X.; Behpour, S.; and Xu, M. 2022. Cryo-shift: reducing domain shift in cryo-electron subtomograms with unsupervised domain adaptation and randomization. *Bioinformatics*, 38(4): 977–984.

Chen, H.; Dou, Q.; Yu, L.; Qin, J.; and Heng, P.-A. 2018. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. *NeuroImage*, 170: 446–455.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848.

Eisenstein, F.; Danev, R.; and Pilhofer, M. 2019. Improved applicability and robustness of fast cryo-electron tomography data acquisition. *Journal of structural biology*, 208(2): 107–114.

Elad, M. 2002. On the origin of the bilateral filter and ways to improve it. *IEEE Transactions on image processing*, 11(10): 1141–1151.

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

Guo, Q.; Lehmer, C.; Martínez-Sánchez, A.; Rudack, T.; Beck, F.; Hartmann, H.; Pérez-Berlanga, M.; Frottin, F.; Hipp, M. S.; Hartl, F. U.; et al. 2018. In situ structure of neuronal C9orf72 poly-GA aggregates reveals proteasome recruitment. *Cell*, 172(4): 696–705.

Hagen, W. J.; Wan, W.; and Briggs, J. A. 2017. Implementation of a cryo-electron tomography tilt-scheme optimized for high resolution subtomogram averaging. *Journal of structural biology*, 197(2): 191–198.

Harar, P.; Herrmann, L.; Grohs, P.; and Haselbach, D. 2023. FakET: Simulating Cryo-Electron Tomograms with Neural Style Transfer. *arXiv preprint arXiv:2304.02011*.

Koning, R. I. 2010. Cryo-electron tomography of cellular microtubules. *Methods in cell biology*, 97: 455–473.

Lehtinen, J.; Munkberg, J.; Hasselgren, J.; Laine, S.; Karras, T.; Aittala, M.; and Aila, T. 2018. Noise2Noise: Learning Image Restoration without Clean Data. In *International Conference on Machine Learning*, 2965–2974. PMLR.

Li, C.; Liu, D.; Li, H.; Zhang, Z.; Lu, G.; Chang, X.; and Cai, W. 2022. Domain adaptive nuclei instance segmentation and classification via category-aware feature alignment and pseudo-labelling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 715–724. Springer.

Liao, X.; Li, W.; Xu, Q.; Wang, X.; Jin, B.; Zhang, X.; Wang, Y.; and Zhang, Y. 2020. Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9394–9402. IEEE.

Liu, D.; Zhang, D.; Song, Y.; Zhang, F.; O'Donnell, L.; Huang, H.; Chen, M.; and Cai, W. 2020. Pdam: A panoptic-level feature alignment framework for unsupervised domain adaptive instance segmentation in microscopy images. *IEEE Transactions on Medical Imaging*, 40(1): 154–165.

Martinez-Sanchez, A.; Lamm, L.; Jasnin, M.; and Phelippeau, H. 2024. Simulating the cellular context in synthetic datasets for cryo-electron tomography. *IEEE Transactions on Medical Imaging*, 1–1.

Oikonomou, C. M.; and Jensen, G. J. 2017. Cellular electron cryotomography: toward structural biology in situ. *Annual review of biochemistry*, 86: 873–896.

O'Reilly, F. J.; Xue, L.; Graziadei, A.; Sinn, L.; Lenz, S.; Tegunov, D.; Blötz, C.; Singh, N.; Hagen, W. J.; Cramer, P.; et al. 2020. In-cell architecture of an actively transcribing-translating expressome. *Science*, 369(6503): 554–557.

Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; and Ferrin, T. E. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13): 1605–1612.

Prichard, A.; Lee, J.; Laughlin, T. G.; Lee, A.; Thomas, K. P.; Sy, A. E.; Spencer, T.; Asavavimol, A.; Cafferata, A.; Cameron, M.; et al. 2023. Identifying the core genome of the nucleus-forming bacteriophage family and characterization of Erwinia phage RAY. *Cell reports*, 42(5).

Shin, H.; Kim, H.; Kim, S.; Jun, Y.; Eo, T.; and Hwang, D. 2023. SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7412–7421. IEEE.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33: 596–608.

Van Vliet, L. J.; Young, I. T.; and Beckers, G. L. 1989. A nonlinear Laplace operator as edge detector in noisy images. *Computer vision, graphics, and image processing*, 45(2): 167–195.

Wan, W.; and Briggs, J. A. 2016. Cryo-electron tomography and subtomogram averaging. *Methods in enzymology*, 579: 329–367.

Wan, W.; Khavnekar, S.; and Wagner, J. 2024. STOPGAP: an open-source package for template matching, subtomogram alignment and classification. *Acta Crystallographica Section D: Structural Biology*, 80(5).

Wang, L.; Zhang, Y.; and Feng, J. 2005. On the Euclidean distance of images. *IEEE transactions on pattern analysis and machine intelligence*, 27(8): 1334–1339.

wwPDB Consortium, T. 2024. EMDB—the electron microscopy data bank. *Nucleic acids research*, 52(D1): D456–D465.

Xian, J.; Li, X.; Tu, D.; Zhu, S.; Zhang, C.; Liu, X.; Li, X.; and Yang, X. 2023. Unsupervised cross-modality adaptation via dual structural-oriented guidance for 3D medical image segmentation. *IEEE Transactions on Medical Imaging*.

Xie, Q.; Li, Y.; He, N.; Ning, M.; Ma, K.; Wang, G.; Lian, Y.; and Zheng, Y. 2022. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*.

Xu, Z.; Gong, H.; Wan, X.; and Li, H. 2023. Asc: Appearance and structure consistency for unsupervised domain adaptation in fetal brain mri segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 325–335. Springer.

Zeng, X.; Kahng, A.; Xue, L.; Mahamid, J.; Chang, Y.-W.; and Xu, M. 2023. High-throughput cryo-ET structural pattern mining by unsupervised deep iterative subtomogram clustering. *Proceedings of the National Academy of Sciences*, 120(15): e2213149120.

Zeng, X.; and Xu, M. 2020. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4073–4084. IEEE.

Zhang, J.; Chao, H.; Dhurandhar, A.; Chen, P.-Y.; Tajer, A.; Xu, Y.; and Yan, P. 2023a. Spectral Adversarial MixUp for Few-Shot Unsupervised Domain Adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 728–738. Springer.

Zhang, X.; Wu, Y.; Angelini, E.; Li, A.; Guo, J.; Rasmussen, J. M.; O'Connor, T. G.; Wadhwa, P. D.; Jackowski, A. P.; Li, H.; et al. 2024. MAPSeg: Unified Unsupervised Domain Adaptation for Heterogeneous Medical Image Segmentation Based on 3D Masked Autoencoding and Pseudo-Labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5851–5862.

Zhang, Y.; Wang, Y.; Xu, L.; Yao, Y.; Qian, W.; and Qi, L. 2023b. ST-GAN: A Swin Transformer-Based Generative Adversarial Network for Unsupervised Domain Adaptation of Cross-Modality Cardiac Segmentation. *IEEE Journal of Biomedical and Health Informatics*.

Zhao, X.; Mithun, N. C.; Rajvanshi, A.; Chiu, H.-P.; and Samarasekera, S. 2024. Unsupervised Domain Adaptation for Semantic Segmentation with Pseudo Label Self-Refinement. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2399–2409. IEEE.

Zhao, Z.; Zhou, F.; Xu, K.; Zeng, Z.; Guan, C.; and Zhou, S. K. 2022. LE-UDA: Label-efficient unsupervised domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(3): 633–646.

Zheng, X.; Zhu, J.; Liu, Y.; Cao, Z.; Fu, C.; and Wang, L. 2023. Both style and distortion matter: Dual-path unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1285–1295. IEEE.

Zhou, B.; Yu, H.; Zeng, X.; Yang, X.; Zhang, J.; and Xu, M. 2021. One-shot learning with attention-guided segmentation in cryo-electron tomography. *Frontiers in Molecular Biosciences*, 7: 613347.

Zhu, X.; Chen, J.; Zeng, X.; Liang, J.; Li, C.; Liu, S.; Behpour, S.; and Xu, M. 2021. Weakly supervised 3d semantic segmentation using cross-image consensus and inter-voxel affinity relations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2834–2844. IEEE.

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