

Towards Source-Free Cross Tissues Histopathological Cell Segmentation via Target-Specific Finetuning

Zhongyu Li*, Chaoqun Li*, Xiangde Luo*, Yitian Zhou, Jihua Zhu, Cunbao Xu, Meng Yang, Yenan Wu, Yifeng Chen**

Abstract—Recognition and quantitative analytics of histopathological cells are the golden standard for diagnosing multiple cancers. Despite recent advances in deep learning techniques that have been widely investigated for the automated segmentation of various types of histopathological cells, the heavy dependency on specific histopathological image types with sufficient supervised annotations, as well as the limited access to clinical data in hospitals, still pose significant challenges in the application of computer-aided diagnosis in pathology. In this paper, we focus on the model generalization of cell segmentation towards cross-tissue histopathological images. Remarkably, a novel target-specific finetuning-based self-supervised domain adaptation framework is proposed to transfer the cell segmentation model to unlabeled target datasets, without access to source datasets and annotations. When performed on the target unlabeled histopathological image set, the proposed method only needs to tune very few parameters of the pre-trained model in a self-supervised manner. Considering the morphological properties of pathological cells, we introduce two constraint terms at both local and global levels into this framework to access more reliable predictions. The proposed cross-domain framework is validated on three different types of histopathological tissues, showing promising performance in self-supervised cell segmentation. Additionally, the whole framework can be further applied to clinical tools in pathology without accessing the original training image data. The code and dataset are released at: <https://github.com/NeuronXJTU/SFDA-CellSeg>.

Index Terms—Cell Segmentation, Domain Adaptation, Consistency Regularization, Self-supervised Learning.

I. INTRODUCTION

Pathological images are treated as golden standards for clinical diagnosis and cancer grading, where doctors can carefully examine morphologies and appearances from cells to whole-slice tissues for quantitative and qualitative evaluation. Especially, recognition and quantitative analysis of cells play critical roles in pathological examinations, which can help doctors to identify particular subtypes of cancers, evaluate cancer stages, and correlate with genetic mutations. In clinical applications, analytics of cells generally include the identification of cell types, description of cell morphologies, computation of the percentage of nucleus and cells, etc. Traditionally, these works like cell segmentation, detection, and identification are all finished by pathologists and doctors, which are extremely labor-intensive and time-consuming. As reported in Metter et al. [1], the number of pathologists is far from enough in most countries.

To alleviate these problems, deep learning techniques have been investigated for computer-aided diagnosis of pathological images in recent years. Notably, in the task of cell/nuclei segmentation, Janssens et al. [2] exploited supervised learning techniques to segment nuclei in follicular lymphoma pathological images by using a K-NN [3] classifier to learn a mapping from the traditional RGB space to the MDC. And lately, Chen et al. [4] proposed a deep contour-aware network (DCAN) to establish better cell segmentation by a multi-task learning framework that learns probability maps and clear contours in a single network at the same time. As for cell segmentation on different kinds of tissues, Haq et al. [5] proposed a self-supervised domain adaptation framework based on GAN [6] to do domain adaptation between Kidney Renal Clear cell carcinoma (KIRC) and Triple Negative Breast Cancer (TNBC) tissues and got good segmentation results on both domains. Besides, there are also multiple public challenges and datasets have been released for the task of cell segmentation, including the data science bowl [7], MoNuSeg (Multi-organ nuclei segmentation challenge) [8] and MoNuSAC (Multi-organ nuclei segmentation and classification challenge) [9]. To train these fully-supervised methods above, sufficient annotations with specific types of images are always required, where images have corresponding pixel-level ground-truth labels [10]. Accordingly, many researchers made public datasets to help other

This work was supported in part by the National Key Research and Development Program of China (2020YFB1711500), the National Natural Science Foundation of China (61902310), and the Natural Science Basic Research Program of Shaanxi (2020JQ-030).

Z. Li, C. Li and X. Luo contributed equally. Corresponding author: Y. Chen (Email: chenyifeng716@qq.com)

Z. Li, C. Li, Y. Zhou and J. Zhu are with the School of Software Engineering, Xi'an Jiaotong University, Xi'an, China.

X. Luo is with School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China.

C. Xu and Y. Chen are with the Department of Pathology, Quanzhou First Hospital Affiliated to Fujian Medical University, Quanzhou, China.

M. Yang and Y. Wu are with the Frontline Intelligent Technology (Nanjing) Co., Ltd, Nanjing, China.

researchers to do research on this task, like the KIRC dataset [11] which consists of Kidney Renal Clear cell carcinoma (KIRC) pathological images and labels, Triple Negative Breast Cancer Cell (TNBC) pathological images and labels data [12], TCIA dataset [13] that consists pathological images from 14 different tissues, and Pannuke dataset [14] that consists pathological images from 19 different tissues.

Despite the above novel methods and public repositories having facilitated the research on pathological CAD, there are still multiple challenges that prevent the application of deep learning on clinical cell segmentation and quantification. Firstly, cell segmentation is indeed a challenging task. The background of pathological images is complex and noisy, especially considering the segmentation at the cell level. The cell segmentation task generally requires a lot of training samples with good annotations to support the fully-supervised deep neural network. However, manual annotation of cells is highly labor-intensive and time-consuming. Secondly, most deep models are only suitable for the given training set, which cannot be generalized to other image modalities. However, there are hundreds to thousands of types of tissues and cells in pathological diagnosis. Pathological images and cells generally demonstrate different appearances, which can be imaged from various tissues using different staining methods and microscopes. It is impossible to collect sufficient well-annotated pathological images for all types of tissues and cells. Thirdly, in clinical applications, hospitals, and medical centers generally have separate information systems with strict requirements of data privacy, i.e., medical data cannot be transferred outside. However, most current transfer learning and domain adaptation algorithms need to access the original training data with annotation in the source domain [15]. It's hard to access the original pathological images and transfer their trained model to other pathological image types.

Taking the challenges above into account, this paper proposes an effective and simple framework for cross-domain segmentation of cells towards different tissues and medical centers. Notably, we first apply a fully-supervised cell segmentation network on the source domain. Then we simply employ the trained fully-supervised model on the pathological images of the target domain for further task-specific fine-tuning. We demonstrate that different layers in the network perform various functions during the domain adaptation of cell segmentation. The framework can achieve excellent performance on the domain adaptation of cell segmentation by fixing and tuning different layers in a self-supervised manner. The contribution of this paper can be summarized in four aspects:

- 1) To the best of our knowledge, this is the first work that focuses on the source-free domain adaptation of pathological cell segmentation, crossing different tissues and multiple medical centers without accessing original source datasets;
- 2) In the framework, we propose a simple but effective solution to adapt the original cell segmentation model on different target-specific pathological tissues, i.e., fine-tuning the pre-trained model on batch normalization layers along with the final classification layer in a self-supervised manner;
- 3) Two constraint terms are introduced in the self-supervised pipeline on both local and global levels to well predict the

edges of pathological cells;

- 4) Extensive and comprehensive experiments are carried out on three pathological datasets from different tissue and cell types, demonstrating effective and superior performance in comparison with other domain adaptation methods.

The following of this paper is organized as follows: Section 2 briefly reviews related works of cell segmentation and domain adaptation. Section 3 provides details of the proposed cell segmentation framework of source-free self-supervised domain adaptation, followed by experimental results and ablation studies in Section 4. Finally, Section 5 concludes the paper and discusses future works.

II. RELATED WORKS

A. Cell Segmentation

Before deep learning methods were used in cell segmentation, there are plenty of traditional segmentation methods [16], [17] widely used in this field, such as global threshold methods [18]–[21], Watershed Transform [22]–[25], clustering [26]–[29], graph-based methods [30]–[33], and pixel-level methods [34]. Nowadays, cell segmentation in histopathology images has been extensively studied with a variety of deep learning methods. Inspired by the development of Fully Convolutional Networks (FCN) [35], a variety of deep learning methods are proposed in the segmentation field, such as U-Net [36], DeepLab [37], and UNet++ [38]. Naturally, many researchers applied these methods to cell segmentation tasks and achieved good results. For example, Janssens et al. [2] applied a multiclass SVM [39] classifier that classifies segments into three categories to cell segmentation on H&E-stained [40] skeletal muscle images.

However, these methods are not specifically modified for the cell segmentation task as they did not consider the difference between cell segmentation and normal semantic segmentation methods. For example, the cells' property of cell adhesion makes it challenging to differentiate and segment a single cell from others. To address this problem, Chen et al. [4] proposed a deep contour-aware network (DCAN) to establish better segmentation by a multi-task learning framework, which learns probability maps along with precise contours in a single network at the same time and combines them to get the final predictions. Sadanandan et al. [41] proposed an approach for the creation of ground truth for the segmentation of bright-field images of cultured cells based on end-point fluorescent staining. They introduced DeLTA (Deep Learning for Time-lapse Analysis), an image processing tool that uses two U-Net deep learning models consecutively to first segment cells in microscopy images and then perform tracking and lineage reconstruction. Lal et al. [42] proposed a method that uses a network formulated as a Triple U-Net [36] structure that includes an RGB branch, a Hematoxylin branch, and a Segmentation branch. Multiple competitions are also held to organize researchers to overcome problems in this field, including the data science bowl in 2018 [7] and MoNuSeg (Multi-organ nuclei segmentation challenge) [8]. Moreover, with the development of segmentation methods, some segmentation tools are also proposed to help non-experts to do segmentation [43].

Despite the above fully supervised methods could indeed do well in many pathological image datasets. However, well-annotated datasets need a lot of labor-intensive labeling work. That's why the amount of cell segmentation datasets is extremely limited in clinical diagnosis. Therefore, the application of these fully supervised methods is restricted. Accordingly, researchers start to work on self-supervised methods for this problem, like Janssens's [2] method. They used a support vector machine (SVM) classifier [39] and a weak edge enhancement operator (WEEO) to achieve segmentation without any label. However, the performance of these methods is far behind the supervised methods.

B. Domain adaptation

Self-supervised Domain Adaptation: Under the hypothesis of the source domain and the target domain having the same feature space, domain adaptation methods have been used in the cell segmentation field. Currently, the existing domain adaptation methods for segmentation can be roughly categorized into adversarial learning-based methods and self-supervised learning-based methods. Liu et al. [15] proposed a method based on Cycada [44] which adds a task re-weighting mechanism along with a nuclei inpainting mechanism to improve the performance of the framework on data from different organs. However, these self-supervised methods' performances still have a long way to go compared with methods that use both labeled and unlabeled data like domain adaptation. Some self-supervised domain adaptation approaches for cell segmentation have been proposed. Vu et al. [45] proposed a method that addresses the task of self-supervised domain adaptation with losses based on the entropy of the pixel-wise predictions and uses entropy loss and adversarial loss, respectively. Haq et al. [5] proposed a framework based on GAN [6] along with a reconstruction network to do domain adaptation in different organs.

Regarding self-supervised-learning-based methods, Yang et al. [46] proposed a method that uses a simple Fourier Transform and its inverse, simplifying the domain adaptation training and achieving good results in semantic segmentation. Even though the above methods have already considered different datasets and obtained well-performed results, there are still multiple challenges in cell segmentation. Usually, domain adaptation methods need full access to the source domain while training, which causes the problem that these methods typically lack data safety and may not satisfy the practical need in actual production.

Semi-supervised Domain adaptation: As additional annotated data is typically used in actual domain adaptation scenarios to improve the reliability of segmentation prediction, semi-supervised are proposed by using limited labels in the target domain [47]–[54]. These methods can achieve better results with a small amount of manual labeling work. For instance, Tarvainen et al. [51] proposed the mean-teacher method, which uses two models updated by using consistency cost between them to learn in a semi-supervised way. Cui et al. [52] applied the mean-teacher method on semi-supervised brain lesion segmentation tasks and achieved good performance. Hung et

al. [55] provided a semi-supervised domain adaptation method that uses GAN [6] method. Additionally, Wang et al. [53] proposed the Alleviating Semantic-level Shift (ASS) method, which augments the synthetic source data with annotated samples from the target domain for semi-supervised domain adaptation. More recently, Li et al. [54] introduced a dual-teacher framework that exploits the intra-domain and inter-domain knowledge for semi-supervised domain adaptation.

Source-free Domain Adaptation: For the domain adaptation scenario when data from source domains are not accessible, researchers proposed methods for source-free domain adaptation. For example, Liu et al. [56] established a source-free domain adaptation method relying on the entropy minimization method and uses the exponential momentum decay scheme and the transferability adaptive high-order statistics consistency loss to improve the methods' performance. Wang et al. [57] proposed a framework that separates the training step on the source and target domain by freezing the network's parameters except for the batch normalization layers. Inspired by Tent [57], Hu et al. [58] extended it to the medical image segmentation task by combining the regional nuclear-norm and contour regularization. However, this method only uses a simple entropy minimization loss without considering the consistency of the target domain. Fu et al. [59] proposed a method that uses regularization constraints for source-free transfer learning on detection and recognition tasks. Chen et al. [60] proposed a method based on denoised pseudo-labeling. By adding supervision to the pseudo label generating method, the method could improve its performance on source-free domain adaptation. Hou et al. [61] proposed a source-free domain adaptive segmentation framework that uses entropic objective loss for implicitly aligning the features and a bi-directional self-training strategy to generate pseudo source domain-style images by learning from the knowledge difference between the source model and the target model. Additionally, Fleuret et al. [62] trained the target domain by using the pseudo images generated by the model trained on the source and using both the prediction and the noise generated by two networks with the same sharing encoder to learn information from. However, none of the source-free domain adaptation methods are specially designed for the segmentation of pathological cells under the practical demands of cross-tissue label limitations and medical center data privacy.

III. METHODOLOGY

A. Overview

The complete framework is shown in Fig. 1. We use a CNN-based segmentation network (e.g., U-Net [36]) as our framework's backbone due to their good performances on existing cell segmentation tasks. We only use the model trained on the source domain and the target images to do domain adaptation by finetuning the batch normalization layers' and final classification layer's parameters by applying self-supervised methods and data augmentation methods.

Formulation We aim to develop a source-data-free domain adaptive cell segmentation with only a pre-trained model from the source domain in this paper. Specifically, we have

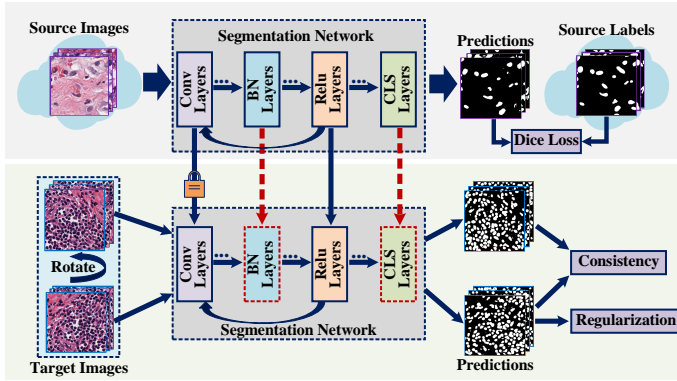


Fig. 1: Overview of the proposed framework towards source-free cross-domain histopathological cell segmentation via target-specific finetuning. Source images are only used for training the fully-supervised segmentation network. In training target images, network parameters are only updated in batch normalization layers and convolutional classification layers, in a self-supervised manner with consistency in both global pathological images and local pixel-level cells.

n_s labeled images $\{x_s^i, y_s^i\}_{i=1}^{n_s}$ from the source domain D_s and n_t unlabeled images $\{x_t^i\}_{i=1}^{n_t}$ from the target domain D_t , where $x_s^i \in X_s$, $y_s^i \in Y_s$, and $x_t^i \in X_t$. And the size of the histopathological image patches is presented as $H \times W \times 3$, while the pixel-wise ground-truth label's size is $H \times W \times 1$. Our goal is to learn a network f_t that could achieve accurate cell segmentation results on target images $\hat{Y}_t = f_t(X_t)$ by domain adaptation from the fully-supervised segmentation network f_s trained on the source domain. In our framework, we only access target images $\{x_t^i\}_{i=1}^{n_t}$ and the trained source segmentation network f_s .

The source segmentation network takes images X_s as input and the segmentation predictions $\hat{Y}_s = f_s(X_s)$ of the same width and height as output. We train the source segmentation network f_s to generate the predictions \hat{Y}_s through a fully-supervised segmentation network while using Y_s as the ground-truth labels of the source domain. For the target domain, we freeze all parameters except for batch normalization layers' and the output convolution layer's in the trained f_s to make the network could be trained in the normalization part and the output convolutional part. We consider it as f_t and use images X_t as the input and $\hat{Y}_t = f_t(X_t)$ as the output.

B. Fully-supervised training on the source domain

In the source domain, we use a traditional fully-supervised method to train the model for the prediction of the histopathological images' pixel-level cell segmentation results. As pathological images generally have smaller foreground than semantic segmentation tasks in nature images, the dice-coefficient loss is more effective than the binary cross-entropy loss in segmentation tasks on pathological images. Accordingly, we use the dice-coefficient loss as our segmentation loss:

$$L_{\text{dice}} = 1 - \frac{2Y'_s \cdot \hat{Y}'_s}{Y'_s + \hat{Y}'_s} \quad (1)$$

where Y'_s and \hat{Y}'_s are flatten Y_s and \hat{Y}_s respectively. After training the network, we only use the trained deep model in further operation to guarantee data privacy in the source domain.

C. Domain adaptation with target-specific finetuning

When we put the target images into the network, the output may not good as the fact that the source and target domain have different domain centers. And what we need to do in this part is precisely move the target domain center and make it closer to the source domain. Here, we develop a self-supervised method to finetune the model.

Entropy and nuclear-norm for cell segmentation Since there is a domain offset between the source domain and the target domain, the parameters of the backbone need to be fine-tuned to fit the target domain. Entropy minimization is widely used in semi-supervised learning to utilize the unlabeled data [63]. In this part, the entropy of the output \hat{Y}_t is minimized, as in previous semi-supervised and domain adaptation methods. By doing so, the results could have smaller entropy, which is shown as the pixels are clustered to the two sides of the prediction. This encourages the model to learn from unlabeled pixels by min-max the intra/inter-class discrepancy, maximizing the difference of the inter-class and making the intra-class prediction results more confident.

As normalization could scale and shift the features' distribution while not changing the convolution kernels that extract the feature, for the network's stability and efficiency, only parameters in the batch normalization layers are updated to alleviate the domain adaptation problem without accessing source labels. The network's output classification layer is also updated as it may be more sensitive to the changes in the input images' distribution. According to Fig.1, since many layers (e.g., later convolutional layers) include high-dimensional features for pixel-level segmentation, the fine-tuned deep models are generally more sensitive to these high-dimensional features when transferring from source to target domain, which may lead to bias for pixel-level segmentation. On the other hand, batch-norm layers tune model parameters in linear and low-dimensional space. In each convolutional module, including a convolutional layer, a batch-norm layer, and a ReLU layer, the output of batch-norm layers is set as the input of the next layer's convolution. This strategy gradually fine-tunes model parameters without affecting high-dimensional inference. Similarly, the final classification layers tune model parameters by mapping high-dimensional features to low-dimensional prediction matrices.

In practice, we measure the task predictions by using Shannon entropy [64], which could be presented as $H(\hat{Y}_t) = p(\hat{Y}_t) \log p(\hat{Y}_t)$ for our two categories segmentation task. As a measure of the task, $H(\hat{Y}_t = f_t(X_t))$ is a function of the network's parameters and we could minimize it by fine-tuning it as a loss function as follows:

$$L_{em} = -\frac{1}{H * W} \sum_{h=0}^H \sum_{w=0}^W \hat{Y}_t \log(\hat{Y}_t) \quad (2)$$

Where H and W are the predictions' sizes.

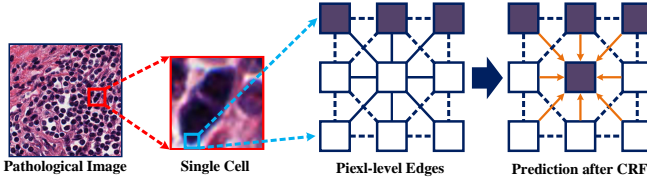


Fig. 2: We use the conditional random field as a learnable loss to maintain the results' local consistency. From left to right: a pathological image patch, a single cell image, a 3×3 pixel matrix with different intensities, pixel-level prediction of the center pixel based on surrounding 8 pixels.

Although the entropy minimization loss has adjusted the model to fit the target domain, it still doesn't consider the information that can be directly extracted from the target domain. Accordingly, we introduce batch nuclear-norm maximization method [65] which maximizes nuclear-norm to improve both discriminability and diversity of the prediction. The mathematical formulation is as follows:

$$L_{bnm} = -||S(f_t(X_t))||_* \quad (3)$$

Where S presents the singular value matrix, $|| \cdot ||_*$ presents the nuclear-norm calculation. By applying the entropy minimization and the diversity, the prediction distribution can be more specific. And the results may be more confident for segmentation on the target domain.

Global and local consistency for cell segmentation We use entropy and nuclear-norm to measure our predictions. However, our framework still lacks consistency as the entropy and nuclear-norm only learn from pixel-level information and couldn't put the problem into the whole image. According to former methods in unsupervised learning [48] [51] [66], the data augmentation methods like the rotation method could guide the framework's global consistency without using labels, we introduce auxiliary rotation loss to guide the framework's global consistency. This loss is based on the rotation consistency, i.e., the prediction of the rotated images should be consistent with the rotated prediction of the original images. However, this auxiliary rotation loss still cannot alleviate the problem that mini gaps or small fake positive regions for this region may exist whether rotated or not. Then we introduce conditional random field (CRF) loss to our framework to restrain the predictions' local consistency, which could consider the pixels' surrounding areas' prediction to give predictions. In this part, we minimize the mean square error between the rotated prediction $\hat{Y}_{t,r}$ and the origin prediction \hat{T}_t as our auxiliary rotation loss as follows:

$$L_{rot} = ||\hat{Y}_{t,r} - \hat{Y}_t||^2 \quad (4)$$

Where r presents the rotation operation, which could also be replaced with other data augmentation methods.

As cells in whole slide images always have different styles from tissue fluid around them and relatively have local similarity, we can use this local consistency to help the domain adaptation. In this part, we employ a gated conditional random

field loss [67] to minimize the performance gap between the original target images and the cell segmentation prediction by using the same nature of both the source and the target domain. The former usages of this gated CRF loss function are mainly employed on weakly supervised domain adaptation tasks for extracting knowledge learned from weak annotation to accurate spacial level-related pixel predictions. Unlike the above common usages, we employ this method in a source-free domain adaptation task for better adjusting the prediction results. According to the property of CRF in learning local spatial relativity [68], every spatial level-related pixel (e.g., each pixel of a 3×3 matrix in Fig.2) is weighted for pixel-level segmentation predictions. Especially, this strategy can better alleviate predictions of corners and outliers when tackling the gap between the source and target domain. Meanwhile, the CRF loss could also learn from the images' local consistency and prevent overfitting [69] of the "fake" label methods. Fig. 2 shows how it works. To employ the CRF in a loss function, we first define the energy term of labeling classes i and j at positions a and b , as:

$$\psi_{a,b(i,j)} = \mu(i,j)K_{ab} \quad (5)$$

$$K_{ab} = \sum_{p=1}^P w^{(p)} k^{(p)}(f_a^{(p)} f_b^{(p)}) \quad (6)$$

Where $\psi_{a,b}$ denotes the pairwise potentials, and μ is a generic class compatibility matrix of size $C \times C$, and here C is 1 as we are working on a two categories segmentation. Then here follows a mixture of P kernels $k^{(p)}(\cdot, \cdot)$ for a pair of positions with weights $w^{(p)}$. The feature vector $f_a^{(p)}$ is specific to the p -th kernel and doesn't depend on the prediction. And in practice, it is made up of position coordinates or input modalities like RGB or depth. Then we use Gaussian kernels with kernel-specific bandwidth parameters $\sigma^{(p)}$ and Pott's class compatibility model:

$$k^{(p)}(f_a^{(p)}, f_b^{(p)}) = \exp\{-\frac{1}{2}|\frac{f_a^{(p)} - f_b^{(p)}}{\sigma^{(p)}}|^2\} \quad (7)$$

$$\mu(i,j) = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{otherwise.} \end{cases} \quad (8)$$

We can see that given the feature-wise similarity, and if the labeling of classes at two positions is different, its value is higher. Meanwhile, the color of the pixel should also contribute to the outputs. We use pixels' distance and color to define $f_a^{(p)}$ and $f_b^{(p)}$. And We can write the energy of prediction \hat{Y}_t at positions a and b , and the total energy of the prediction \hat{Y}_t , where every pixel of the image is related to each other:

$$\psi_{a,b}(\hat{Y}_t) = \sum_{i,j \in [0,C]} \psi_{a,b(i,j)} \hat{Y}_{t,a}(i,j) \hat{Y}_{t,a}(i) \hat{Y}_{t,b}(j) \quad (9)$$

$$\Psi(\hat{Y}_t) = \sum_{a=1}^N \sum_{b=1 \neq a}^N \psi_{a,b}(\hat{Y}_t) \quad (10)$$

In $\hat{Y}_{t,a}(i,j) \hat{Y}_{t,a}(i) \hat{Y}_{t,b}(j)$, the predictions represent that the selected points a and b 's outputs are i and j respectively. These

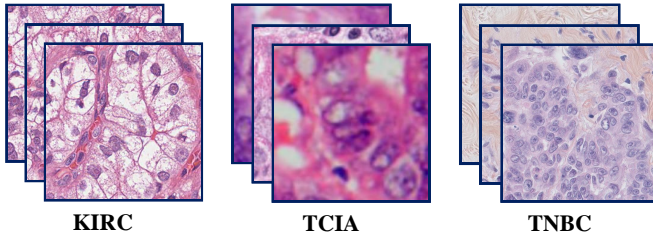


Fig. 3: Sample images of three datasets.

three items should be 1 when a and b are selected. For any position a, if we assume that all kernels are Gaussian and feature vectors $k^{(p)}$ contain position coordinates, the pairwise energy of jointly labeling a and b would decrease with the distance from a. Meanwhile, a CRF loss with a long distance from a to b may bring great computational challenges to the task and may not be beneficial to constrain the local consistency. Thus, we limit the range of summation of b to a local neighborhood $\Omega(a)$ characterized by a function of the acceptable loss of signal ϵ and kernel bandwidth $\sigma^{(p)}$. We define the two-dimensional local neighborhood window with radius r of the position $a(a_x, a_y)$ (a_x and a_y are a's coordinates) as $\Omega_r(a) = [a_x - r, a_x + r] \times [a_y - r, a_y + r]$. Then the total energy becomes:

$$L_{crf} = \Psi(\hat{Y}_t) = \sum_{a=1}^N \sum_{b \in \Omega_r(a)} \psi_{a,b}(\hat{Y}_t) \quad (11)$$

By the above loss function, the CRF is embedded into the network, which makes the local influences propagate to the whole image while training.

Overall, we optimize the following total loss when training our framework on the target domain:

$$Loss_{target} = L_{em} + L_{bnm} + \phi_{rot} L_{rot} + \phi_{crf} L_{crf} \quad (12)$$

Where $\phi_{rot} = k * e^{-(i+E)^2}$, in which i presents the epoch, k scales the maximum value of the weighting function, and E defines when the weight comes to the peak. In our experiments, we set k to 1.0 and E to 10 based on experience. And $\phi_{crf} = \alpha i$, if $\phi_{crf} < 1.0$, in which i presents the epoch, α presents the growth speed of the weight. We set α to 0.05 based on our experiences.

IV. EXPERIMENTS

A. Datasets

We use three pathological image datasets in our experiments. The KIRC dataset was released by Irshad et al. [11] and consists of 463 images of 400×400 pixel size. Pathological cells were labeled by pathologists and came from whole slide images (WSI) of Kidney Renal Clear cell carcinoma (KIRC). Naylor et al. [12] released a dataset of 50 images of 512×512 pixel size that came from whole slides of Triple-Negative Breast Cancer Cell (TNBC). We refer to it as the TNBC dataset here. Hou et al. [13] released the TCIA dataset that consists of 1356 images of 256×256 pixel size from slides

TABLE I: List of number, size, and scale of images of three datasets, i.e., TCIA [13], TNBC [12], and KIRC [11].

Datasets	TCIA	TNBC	KIRC
Image number	97	50	462
Size	256×256	512×512	400×400
Scale	$40 \times$	$40 \times$	$40 \times$

of 14 different cancer types. We only use 97 images from Stomach adenocarcinoma (STAD) in the TCIA dataset for our experiments, focusing on multi-center domain adaptation across different tissues. To simplify the illustration, we denote the source domain dataset A and the target domain dataset B as A->B.

As presented in Fig. 3, images in these three datasets demonstrate significant domain gaps. The TNBC dataset's stain style is quite different from the other two datasets, indicating stain differences among these datasets. We also observe that the images in TCIA are slightly more blurred than those in the other two datasets, indicating differences in image clarity. Furthermore, we notice that the tissues around the cells differ significantly in shape and color. Moreover, since the images come from different tissues in different datasets, there is a clear domain gap between the information in each dataset.

B. Implement Details

1) *Settings*: In our framework, we use U-Net [36] as our segmentation network. We use 80% of each dataset's images for training, 10% for validation, and 10% for evaluation. We train on different domain adaptation combinations of TNBC->KIRC, TCIA->KIRC, TNBC->TCIA, and TCIA->TNBC, which enables us to have a comprehensive evaluation of the performance on different domain adaptation situations. During the training, we resize all the input images to 512×512 . We employ Adam optimizer [70] to optimize the losses with learning rates of 0.001, a weight decay of 0.001, a batch size of 8 for source data, and 4 for the target data when training on the source domain and the target domain, respectively. As for the CRF loss function, we choose a kernel size of 5×5 to get the best performance for the cell segmentation task. We train 400 epochs on the source domain and 200 epochs on the target domain. And the network always reaches its best performance within 50 epochs on the target domain. All experiments are carried out by using the Pytorch framework on a Linux system with 2 RTX2080Ti graphic cards and take about 16 GB of memory running for 8 hours.

2) *Evaluation Metrics*: We follow previous work to report multiple metrics which contain the Dice metric to verify the prediction's accuracy performance. Meanwhile, we also use 95% Hausdorff Distance (HD95) and Average Symmetric Surface Distance (ASSD) [71] to evaluate the shape difference between the ground truth and the prediction. Unlike the Dice metric, the lower of the HD95 and ASSD metrics indicate a more minor difference between the ground truth and the prediction. What's more, we also use the AJI metric to verify the prediction's instance-level performance. Moreover, we report results from all the above domain adaptation settings

TABLE II: Quantitative results of deep models when only finetune parameters in specific layers parameters, using the TNBC [12] dataset as the source domain.

	KIRC			TCIA		
	Dice	HD95	ASSD	Dice	HD95	ASSD
U-Net	0.691	53.877	9.535	0.611	65.720	11.480
Classification convolution	0.699	54.018	9.446	0.627	65.010	10.871
Input convolution	0.622	57.442	10.237	0.610	66.421	11.021
Down1 convolution	0.671	58.265	10.341	0.590	70.271	12.374
Down2 convolution	0.680	54.673	10.447	0.604	68.874	12.116
Down3 convolution	0.656	61.223	10.871	0.601	68.463	12.095
Down4 convolution	0.653	61.341	10.903	0.600	69.014	12.317
Up1 convolution	0.649	61.491	10.985	0.597	70.122	12.285
Up2 convolution	0.607	64.718	11.203	0.538	78.428	14.413
Up3 convolution	0.565	70.727	12.559	0.593	71.380	12.556
Up4 convolution	0.440	83.719	23.016	0.492	81.271	22.291
Batch Normalization	0.705	71.601	13.229	0.633	49.611	9.512
all convolution	0.399	103.247	28.727	0.373	107.394	28.991
all layers	0.408	98.113	26.916	0.418	96.472	26.588

respectively to find out the difference in the performance of these methods. We calculate both the variance and the average of the metrics to get a comprehensive understanding of the performance.

C. Ablation Study

To figure out which layer's parameters have a positive effect on target domain adaptation, we finetune different layers' parameters in the backbone with only an entropy minimization loss function. Accordingly, we presented Table II for testing our backbone's different structures' performances to prove our finetuning setting when applying domain adaptation. All experiments use the same backbone trained on the source domain (TNBC dataset) and don't use any target label while implementing domain adaptation. We only finetune one layer's parameters in a single experiment to control variables and use U-Net as our backbone.

As is shown in Table II, the performances are significantly better than others when finetuning on the Batch Normalization layers and the final classification convolution layer. When finetuning on the convolution layers, the results of the framework turned out to be even worse than straightly applying the model trained on the source domain to the target domain without domain adaptation. Besides, it turns out that the metrics drop down fast to zero when only finetuning convolution layers' parameters. Accordingly, we set our framework by only finetuning the parameters of the Batch Normalization layers and final classification convolution layer when applying domain adaptation.

After comparing our framework with other domain adaptation methods, we have a clear vision of our framework's performance. But we still lack a clear understanding of the contribution of our framework's different single parts. Accordingly, we did ablation experiments by deleting every component from our framework to determine if every component could have contributed to the whole framework. And the results show that with every single component cut from our framework, the performance drops below our framework. Especially when deleting the CRF [67] or the auxiliary rotation loss, the performance of Dice, HD95, ASSD, and AJI all drops a lot due to the lack of consistency constraint, which further

TABLE III: The metric results for ablation experiments in domain adaptation from TNBC to TCIA dataset.

	Dice	HD95	ASSD
U-Net(lower bound)	0.611±0.091	65.720±63.181	11.480±9.866
U-Net(upper bound)	0.874±0.072	20.324±9.334	4.498±1.531
Tent [57]	0.633±0.085	49.611±36.473	9.512±7.186
Ours(without cls)	0.640±0.090	50.571±38.982	9.753±7.798
Ours(without BNM)	0.627±0.095	53.583±38.695	10.152±7.264
Ours(without CRF)	0.626±0.092	51.256±38.626	9.970±7.576
Ours(without rotation)	0.623±0.098	53.929±37.956	9.808±6.439
Ours(only cls)	0.624±0.096	50.118±37.771	9.407±6.398
Ours(only BNM)	0.623±0.088	49.552±38.481	9.448 ±6.539
Ours(only CRF)	0.620±0.102	52.104±37.956	9.526±7.104
Ours(only rotation)	0.618±0.094	51.864±37.766	10.048±6.699
Ours	0.645±0.086	48.415±38.373	9.213±7.433
ResNet34 [72]	0.667±0.087	68.918±65.775	12.422±10.900
ResNet34+Ours	0.732±0.089	62.378±67.097	11.580±10.392

TABLE IV: The metric results for ablation experiments in domain adaptation from TNBC to KIRC dataset.

	Dice	HD95	ASSD
U-Net(lower bound)	0.691±0.053	52.362±22.645	9.535±2.943
U-Net(upper bound)	0.857±0.075	22.135±10.244	4.918±2.387
Tent [57]	0.705±0.101	71.601±78.571	13.229±13.482
Ours(without cls)	0.697±0.098	75.538±86.389	13.407±13.994
Ours(without BNM)	0.730±0.007	53.877±59.764	9.235±7.785
Ours(without CRF)	0.730±0.071	53.767±59.666	9.232±7.791
Ours(without rotation)	0.727±0.072	49.129±54.298	8.920±7.386
Ours(only cls)	0.713±0.083	67.634±53.241	12.391±8.358
Ours(only BNM)	0.722±0.077	58.220±47.381	9.487±4.143
Ours(only CRF)	0.715±0.101	67.226±46.579	12.110±7.104
Ours(only rotation)	0.711±0.095	65.147±57.465	11.743±8.347
Ours	0.733±0.064	41.390±50.404	7.958±6.771
ResNet34 [72]	0.696±0.050	56.203±21.804	9.973±3.179
ResNet34+Ours	0.702±0.052	64.180±27.553	11.124±3.748

TABLE V: The metric results for ablation experiments in domain adaptation from TCIA to TNBC dataset.

	Dice	HD95	ASSD
U-Net(lower bound)	0.351±0.214	115.062±125.286	29.280±41.377
U-Net(upper bound)	0.869±0.073	20.719±9.782	4.671±1.891
Tent [57]	0.686±0.055	27.079±17.928	5.673±2.152
Ours(without cls)	0.686±0.053	26.849±17.962	5.664±2.138
Ours(without BNM)	0.692±0.044	25.671±17.771	5.503±1.976
Ours(without CRF)	0.692±0.045	25.644±17.763	5.499±1.976
Ours(without rotation)	0.689±0.044	25.653±17.782	5.503±1.976
Ours(only cls)	0.685±0.055	26.751±18.193	5.271±2.129
Ours(only BNM)	0.686±0.057	27.044±17.874	5.267±2.031
Ours(only CRF)	0.684±0.047	27.179±17.251	5.246±2.116
Ours(only rotation)	0.688±0.049	26.839±17.557	5.288±1.976
Ours	0.693±0.045	25.649±17.271	5.488±1.969
ResNet34 [72]	0.311±0.144	117.432±89.858	22.590±15.630
ResNet34+Ours	0.622±0.065	38.169±28.951	8.160±5.075

TABLE VI: The metric results for ablation experiments in domain adaptation from TCIA to KIRC dataset.

	Dice	HD95	ASSD
U-Net(lower bound)	0.682±0.051	25.653±15.947	6.076±1.999
U-Net(upper bound)	0.857±0.075	22.135±10.244	4.918±2.387
Tent [57]	0.689±0.051	56.785±23.093	10.144±3.251
Ours(without cls)	0.714±0.053	39.892±21.947	7.945±2.862
Ours(without BNM)	0.713±0.053	40.132±21.692	7.932±2.834
Ours(without CRF)	0.714±0.053	39.904±21.911	7.940±2.858
Ours(without rotation)	0.720±0.059	39.774±22.713	7.954±3.228
Ours(only cls)	0.694±0.059	54.277±22.776	9.775±3.210
Ours(only BNM)	0.703±0.057	46.912±23.173	8.641±2.937
Ours(only CRF)	0.692±0.061	55.387±21.995	9.929±3.112
Ours(only rotation)	0.694±0.066	54.446±22.120	9.714±2.861
Ours	0.740±0.062	35.736±20.700	7.147±2.771
ResNet34 [72]	0.608±0.066	32.202±7.794	9.988±2.565
ResNet34+Ours	0.657±0.056	28.251±17.557	6.315±2.325

proves our local and global consistency's efficiency on source-free cell segmentation tasks across different tissues.

From Table III to Table VI, we provided the ablation study in validating the effectiveness of each proposed component, i.e., the original Tent method [57], our framework without and with only fine-tuning the final classification layer, fine-tuning the BN layers, gated CRF module and rotation consistency module. Besides, we also provide the results of lower and upper bound in the domain adaptation, i.e., predicting target data directly using the model trained on source data, and predicting target data using the model trained on target data. As shown in these tables, all the experiments have improved over the baseline, where our framework turns out to work better than those experiments that delete a single component from ours. These results could prove that every part in our framework could contribute to the final performance and has an improvement compared with the Tent [57] method, which is set as our baseline. Meanwhile, we also did experiments on the situation that only apply one loss function among BNM, CRF, and rotation loss, which could help to validate the effectiveness of every single component. BNM could assist the widely used Entropy Minimization method to get more specific self-supervised restraint. Meanwhile, the CRF and rotation component could contribute to more reliable knowledge from the target domain data as long as limited domain gaps, leading to more effective knowledge learned from the source domain. In our cell segmentation task, the condition is satisfied as it turns out that with every single component added to the framework, our framework could get enhanced performance compared with the baseline.

Contributing to the cell segmentation in different source-free domain adaptation cases (Table III, Table IV, Table V, Table VI), the introduced CRF component also validates the effectiveness of the local consistency on the cell segmentation task. Accordingly, unlike most existing works that introduce global constraints in reducing the domain gap, the local constraint could also contribute to the domain adaptive segmentation task, which is rarely discussed before.

However, we can also find that some results from our framework are not the best when evaluated in ASSD and HD95 metrics (like in TNBC->KIRC) as the results are very close to each other. And our framework also does well when using ResNet34 [72] as the backbone, which results in better metric results than straight using the trained backbone to do segmentation. Accordingly, our framework shows a stable performance on different backbones.

D. Comparison with State-of-the-Art Methods

Besides the ablation study, the comparison of the proposed framework with other state-of-the-art methods is also required, especially with other source-free methods. Accordingly, we compare the performance with other related methods including 3 UDA methods and 4 source-free methods. We proposed Table VII, Table VIII, Table IX, and Table X for the performance of other methods compared with ours in different domains.

1) Compared with Self-supervised Domain Adaptation Methods: In this part, we validate our framework and other self-

supervised domain adaptation methods, which have been presented in recent years and perform well in segmentation tasks. Results from CBST, CellSeg, Tent, Advent, and UncertainDA [62] methods on cross-tissue cell segmentation domain adaptation task are provided for the evaluation as follows:

CBST [73] is a widely used class-balanced self-training domain adaptation framework accomplished by generating pseudo labels according to the prediction of the target images. This method needs full access to the source domain during the domain adaptation process and performs badly on multi-tissue cell segmentation tasks.

CellSeg [5] is another domain adaptation method that could not protect the source data privacy as it uses a generative adversarial network along with a reconstruction network to make the target domain outputs consistent to both the source domain and also the target images. And it achieves a better performance on cell segmentation problems than traditional domain adaptation methods that only use GAN [6] framework.

Advent [45] also doesn't consider the data privacy problem and uses a generative adversarial network. However, it uses the discriminator differently from CellSeg as it uses a backbone with two branches that do the same task. It uses one discriminator to discriminate both branches' outputs to bring the two branches' supervision into the whole framework.

2) Compared with Source-Free Domain Adaptation Methods: Meanwhile, as our framework is based on a source-free method, it is reasonable to compare other state-of-the-art source-free methods to the proposed framework. Accordingly, we choose four source-free methods to do the comparison as follows:

UncertainDA [62] is a method using only a model trained on the source domain while doing the domain adaptation, which means that they could prevent source data privacy problems. It uses two networks with the same sharing encoder and different decoders to generate both the prediction and the noise. When doing domain adaptation, the model trained on the source domain is used to generate pseudo labels of target images, which are used to restrict the target's prediction and the noise generated by another decoder.

Tent [57] method straightly uses the model trained on the source domain and freezes its parameters of all layers except batch normalization layers to train it with entropy minimization loss on the target domain, by which the method could do domain adaptation without donating the source data.

SFDA [60] method is based on denoised pseudo-labeling. It uses two complementary pixel-level and class-level denoising schemes to provide more discriminative and less noisy supervision for source-free model adaptation tasks. By adding these supervisions to the pseudo generating method, the method could efficiently do source-free domain adaptation.

ADOS [56] method establishes source-free domain adaptation relying on a pre-trained segmentation model with BN, which is also based on the entropy minimization method. To improve the methods' performance, an exponential momentum decay scheme and a transferability adaptive high-order statistics consistency loss are added to the framework.

In Table VII, Table VIII, Table IX, and Table X, we show the results of all the segmentation methods above in different

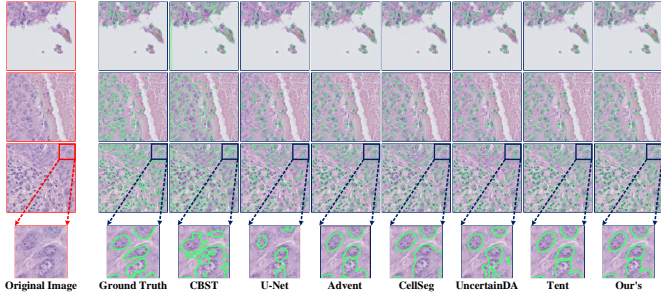


Fig. 4: Visualization results for segmentation from TCIA to TNBC. The top three rows are different images randomly selected from the target domain. From left to right are the results of different comparison methods. We also presents enlarged regions of the third row in the bottom row.

TABLE VII: Quantitative evaluation of different methods in domain adaptation from TNBC to TCIA dataset.

Method	Use source data	Dice	HD95	ASSD	AJI
U-Net [36]	No	0.611±0.090	65.720±63.181	11.480±9.866	0.478±0.064
CBST [73]	Yes	0.725±0.083	29.432±23.711	6.921±3.285	0.617±0.057
CellSeg [5]	Yes	0.744±0.077	63.694±73.776	10.645±9.405	0.638±0.055
AdvEnt [45]	Yes	0.729±0.042	61.433±65.835	10.282±8.547	0.630±0.071
UncertainDA [62]	No	0.669±0.060	63.910±62.295	11.460±9.337	0.538±0.061
Tent [57]	No	0.633±0.085	50.571±38.982	9.753±7.798	0.542±0.062
SFDA [60]	No	0.671±0.051	54.387±42.314	10.181±8.719	0.557±0.063
ADOS [56]	No	0.638±0.064	51.257±39.741	9.832±7.392	0.552±0.061
Ours	No	0.645±0.090	49.611±36.473	9.512±7.186	0.561±0.063

datasets settings and specify whether a method uses source data or not. We also show the predictions of different methods on the domain adaptation of TCIA->TNBC in Fig. 4, where we can see that our framework's prediction is better than others as it has a more precise segmentation and a contour closer to the ground truth in partial regions.

TNBC->TCIA: As it is shown in Table VII, when using the TNBC dataset as the source domain and the TCIA dataset as the target domain, our framework reaches the best performance in the experiments that do not use source data while using U-Net or ResNet34 as the backbone. However, our framework still cannot reach the level of methods that use source data when using U-Net as the backbone.

TCIA->TNBC: As shown in Table VIII, when reversing the datasets by using the TCIA dataset as the source domain and the TNBC dataset as the target domain, our framework reaches the best performance metrics in all source-free methods and among all methods except the CellSeg method, which has full access to the source data.

TNBC->KIRC: When we use the TNBC dataset as the source domain and the KIRC dataset as the target domain, our framework reaches the best performance in all metrics among

TABLE VIII: The result of different methods in domain adaptation from TCIA to TNBC dataset.

Method	Use source data	Dice	HD95	ASSD	AJI
U-Net [36]	No	0.351±0.214	115.062±125.286	29.280±41.377	0.201±0.076
CBST [73]	Yes	0.496±0.146	87.799±84.861	22.590±15.630	0.357±0.071
CellSeg [5]	Yes	0.747±0.087	54.474±33.131	11.133±5.277	0.617±0.067
AdvEnt [45]	Yes	0.606±0.089	50.837±38.724	10.208±7.528	0.503±0.068
UncertainDA [62]	No	0.532±0.117	50.210±25.795	9.742±3.395	0.448±0.067
Tent [57]	No	0.686±0.055	27.079±17.928	5.673±2.152	0.604±0.047
SFDA [60]	No	0.696±0.054	26.971±17.711	5.833±2.782	0.619±0.051
ADOS [56]	No	0.689±0.051	26.468±19.641	5.623±3.132	0.622±0.052
Ours	No	0.693±0.045	25.653±17.782	5.503±1.976	0.629±0.057

TABLE IX: Quantitative evaluation of different methods in domain adaptation from TNBC to KIRC dataset.

Method	Use source data	Dice	HD95	ASSD	AJI
U-Net [36]	No	0.691±0.053	53.877±22.645	9.535±2.943	0.548±0.062
CBST [73]	Yes	0.696±0.040	62.476±22.774	11.372±7.772	0.551±0.061
CellSeg [5]	Yes	0.703±0.065	63.050±24.903	10.682±3.522	0.562±0.057
AdvEnt [45]	Yes	0.686±0.062	54.140±17.692	9.952±3.107	0.565±0.061
UncertainDA [62]	No	0.695±0.050	66.877±19.549	12.448±3.432	0.557±0.069
Tent [57]	No	0.705±0.101	71.601±78.571	13.229±13.482	0.571±0.066
SFDA [60]	No	0.719±0.058	52.014±41.071	9.113±5.142	0.587±0.067
ADOS [56]	No	0.710±0.052	66.173±32.442	12.314±3.758	0.593±0.060
Ours	No	0.730±0.007	52.362±59.764	9.235±7.785	0.626±0.056

TABLE X: The result of different methods in domain adaptation from TCIA to KIRC dataset.

Method	Use source data	Dice	HD95	ASSD	AJI
U-Net [36]	No	0.682±0.051	35.736±15.947	7.147±1.999	0.527±0.067
CBST [73]	Yes	0.699±0.057	27.114±17.269	6.262±2.184	0.586±0.055
CellSeg [5]	Yes	0.717±0.058	26.685±16.701	5.907±2.048	0.604±0.054
AdvEnt [45]	Yes	0.687±0.054	27.223±17.905	5.661±2.099	0.588±0.061
UncertainDA [62]	No	0.701±0.056	26.432±16.963	6.213±2.135	0.574±0.057
Tent [57]	No	0.689±0.051	56.785±23.093	10.144±3.251	0.603±0.056
SFDA [60]	No	0.704±0.067	27.386±21.084	6.493±3.171	0.618±0.061
ADOS [56]	No	0.677±0.061	60.410±25.724	11.426±3.334	0.586±0.059
Ours	No	0.740±0.062	25.653±20.700	6.076±2.771	0.663±0.056

all methods, even beat methods that use source data.

TCIA->KIRC: When we use the TCIA dataset as the source domain and the KIRC dataset as the target domain, our framework reaches the best performance again in all metrics among all methods except for the ASSD and HD95 metric lower than some methods.

Accordingly, we can see that all models using domain adaptation methods have a significant improvement over the single backbone (lower bound), as well as lower performance than the fully supervised method trained on target data with labels. Meanwhile, the results of the domain adaptation methods also fall beyond the experiments that directly use the U-Net and target domain data to do segmentation training (upper bound). For instance, our framework could reach a level that could compete with the methods that use the source data. In some data settings, our framework could even beat those methods using the source data a lot in all metrics. These results show that having full access to the source data may not help too much for domain adaptation tasks like this different tissues' multiple center problem. Our framework beats the Tent method and other source-free methods, which shows that the global and local consistency we proposed is useful for source-free cell segmentation problems even if the data is from different tissues and organs. What's more, our framework also performs well when using ResNet34 [72] as our backbone compared to the direct usage of this backbone, which proves that our framework also has a good versatility that could take effect when being implemented on different kinds of backbones. However, we can also find that in the data settings that have a more significant domain shift (like TCIA->TNBC and TNBC->TCIA), our framework always has a worse performance as the information that the model learned is very limited.

V. DISCUSSION

A. Comparison of Different Data Augmentations

In our framework, we use rotation as the global consistency in our loss function, which could be also considered a data augmentation method. There are also lots of other kinds of data augmentation methods that may perform well in solving

TABLE XI: The metric results for experiments that use different kind of data augmentation method in domain adaptation from KIRC to TCIA and TNBC dataset.

Target Domain	TCIA			TNBC		
	DICE	HD95	ASSD	DICE	HD95	ASSD
Crop	0.702±0.055	45.750±42.144	8.714±8.041	0.701±0.051	66.847±63.431	13.115±11.286
Flip	0.699±0.054	47.189±44.691	9.036±8.056	0.697±0.054	69.554±66.506	13.451±11.224
Ours(rotation)	0.709±0.054	44.855±41.314	8.336±7.797	0.710±0.049	67.698±64.425	12.370±11.073

TABLE XII: The metric results for experiments that use multi different kinds of data augmentation methods in domain adaptation from KIRC to TCIA and TNBC dataset.

Target Domain	TCIA			TNBC		
	DICE	HD95	ASSD	DICE	HD95	ASSD
Rotation+Crop	0.682±0.053	49.847±44.314	9.418±8.364	0.693±0.052	70.115±64.737	13.513±10.403
Rotation+Flip	0.697±0.055	46.875±40.917	9.049±7.285	0.694±0.053	71.443±65.871	14.131±12.217
Flip+Crop	0.688±0.053	48.991±45.330	9.335±7.713	0.684±0.051	72.851±68.476	14.485±12.791

the segmentation tasks. To simplify our framework and avoid the overusing of data augmentation in the contribution of the final performance, we use only one kind of data augmentation strategy that has the best performance compared to other strategies.

As shown in Table XI, we compared our global consistency method with other regular data augmentation methods. The results show that the rotation method achieved the best performance among the three augmentation methods that we tested. In this paper, we choose the rotation method as our global consistency to extract information from the original pathological images. Moreover, with the purpose of a better combination of multiple kinds of data augmentation methods, we did further experiments and showed the results in Table XII. Accordingly, compared to the experiments that use only one single data augmentation method, the framework's performance shows a decrease in every situation that uses multiple data augmentation methods. These results indicate that multiple data augmentation methods may conflict with learning knowledge from the target domain.

B. Robustness Analysis

To verify our framework's robustness, we compare it with other source-free methods on the domain adaptation from KIRC to TCIA and TNBC datasets. As shown in Table XIII and Table XIV, our framework stay in the best performance for the domain adaptation from KIRC to TCIA and TNBC dataset, which could prove that our framework works effectively on the domain adaptation for different scenarios. These results mainly benefited from the CNNs framework's stability and our local and global consistency restrain, which enables the framework to learn data properties from the target dataset to achieve superior performance. Meanwhile, the situation in this case that transfers information from a large dataset (e.g., KIRC) to a small one (e.g., TNBC and TCIA) is similar to the practical situation.

C. Sensibility Analysis of Parameters

Inside our framework, some hyper-parameters may influence the framework's performance. To verify the sensitivity of these hyper-parameters, we did experiments on two hyper-parameters that have the most significance in the framework.

TABLE XIII: The metric results for experiments that use different kinds of source-free domain adaptation methods in domain adaptation from KIRC to TCIA dataset.

	DICE	HD95	ASSD	AJI
Tent [57]	0.684±0.051	54.694±44.314	9.725±6.789	0.597±0.059
SFDA [60]	0.676±0.052	52.173±40.829	9.524±5.547	0.583±0.057
ADOS [56]	0.670±0.053	57.139±44.691	10.326±7.397	0.578±0.061
Ours	0.709±0.054	44.855±41.314	8.336±7.797	0.604±0.063

TABLE XIV: The metric results for experiments that use different kinds of source-free domain adaptation methods in domain adaptation from KIRC to TNBC dataset.

	DICE	HD95	ASSD	AJI
Tent [57]	0.694±0.053	71.557±70.824	13.280±11.296	0.603±0.056
SFDA [60]	0.683±0.052	78.415±68.442	14.713±10.681	0.589±0.057
ADOS [56]	0.703±0.057	70.734±67.441	13.110±11.381	0.609±0.061
Ours	0.710±0.049	67.698±64.425	12.371±11.073	0.613±0.054

As shown in Fig. 5, the result's score shows the sensitivity among different values of the hyper-parameters. Accordingly, the results show Gaussian-like distribution, which could prove the sensitivity and properties of our hyper-parameter settings.

D. Potential use cases

Our proposed framework can achieve excellent domain adaptation without accessing data from the source domain. This advantage is especially suitable for clinical applications, where most raw data cannot be accessed among different medical centers due to data privacy issues. Given the pre-trained weights from the source domain data, our framework achieves source-free domain adaptation by fine-tuning specific layers and global and local consistency. Similar to most of the source-free domain adaptation methods [74], our framework is based on a reliable pre-trained model on the source domain, where the domain gap is not too large. In addition to the source-free domain adaptation, some methods focus on extreme conditions that even cannot access the pre-trained model from the source domain, namely unsupervised black-box model domain adaptation [75]. This kind of method may

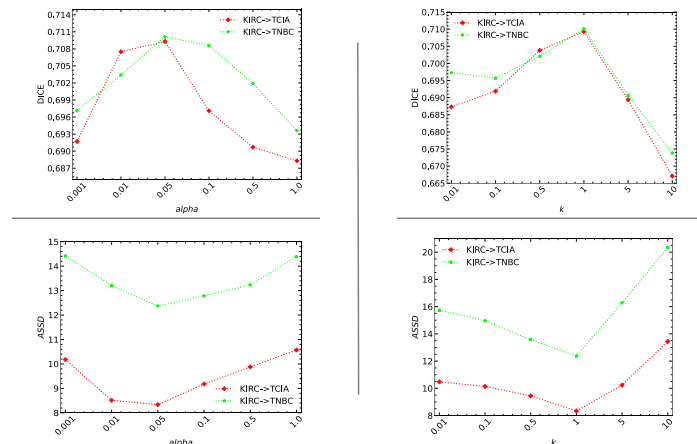


Fig. 5: Sensitivity experimental results of our framework, which include the DICE and ASSD scores under different hyper-parameter settings of α and k .

not perform well on the domain adaptation, especially for the domain adaptive cell segmentation task, where cells from different tissues can be quite different with large domain gaps.

Although the proposed unsupervised framework can achieve excellent performance compared with related domain adaptation methods, there remains a significant gap to the fully supervised methods. In order to merge the cross-domain gap without labor-intensive manual annotation on pixel-level cells, several points can be studied in our future works. Firstly, we may further integrate the current source-free domain adaptation framework with semi-supervised learning, by introducing a few annotated images from the target domain, where the annotations can be either ground truth by users or automatically computed pseudo labels. By introducing few annotations, semi-supervised learning has been validated its superior performance on multiple tasks [51] [52] [54]. Such a semi-supervised manner is especially suitable in medical diagnostic scenarios. Besides, as the diagnostic importance and diversities in histopathological regions and cells can be different, the few manual annotations on the target domain can be automatically selected by active learning, where representative regions/cells are then annotated to boost the semi-supervised domain adaptation framework. In addition, when processing histopathological images from different tissues, types, and medical centers, we found that the color and intensity differences are also noticeable. Accordingly, neural style transfer networks can be further investigated for histopathological images to merge the gap between source and target domains.

VI. CONCLUSION

In this paper, we propose a novel self-supervised domain adaptation method towards a source-free cell segmentation task crossing different pathological tissues. The method does not need to access the source data while implementing the domain adaptation. Moreover, we improve the performance by using global and local consistency in a self-supervised pipeline, which can achieve better discriminability and diversity by maximizing the nuclear-norm of the prediction. Our framework demonstrates superior performance than related state-of-the-arts on the domain adaptation of cell segmentation. In the future, we will investigate source-free domain adaptation methods with semi-supervised learning, improving the transfer learning performance and applying the method to clinical diagnosis.

REFERENCES

- [1] D. M. Metter, T. J. Colgan, S. T. Leung, C. F. Timmons, and J. Y. Park, "Trends in the us and canadian pathologist workforces from 2007 to 2017," *JAMA network open*, vol. 2, no. 5, pp. e194337–e194337, 2019.
- [2] T. Janssens, L. Antanas, S. Derde, I. Vanhorebeek, G. Van den Bergh, and F. G. Grandas, "Charisma: An integrated approach to automatic h&e-stained skeletal muscle cell segmentation using supervised learning and novel robust clump splitting," *Medical image analysis*, vol. 17, no. 8, pp. 1206–1219, 2013.
- [3] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, no. 4, pp. 580–585, 1985.
- [4] H. Chen, X. Qi, L. Yu, and P.-A. Heng, "Dcan: deep contour-aware networks for accurate gland segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2487–2496, 2016.
- [5] M. M. Haq and J. Huang, "Adversarial domain adaptation for cell segmentation," in *Medical Imaging with Deep Learning*, 2020.
- [6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, pp. 2672–2680, 2014.
- [7] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, et al., "Nucleus segmentation across imaging experiments: the 2018 data science bowl," *Nature methods*, vol. 16, no. 12, pp. 1247–1253, 2019.
- [8] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu, et al., "A multi-organ nucleus segmentation challenge," *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1380–1391, 2019.
- [9] R. Verma, N. Kumar, A. Patil, N. C. Kurian, S. Rane, S. Graham, Q. D. Vu, M. Zwager, S. E. A. Raza, N. Rajpoot, et al., "Monusac2020: A multi-organ nuclei segmentation and classification challenge," *IEEE Transactions on Medical Imaging*, 2021.
- [10] N. Kumar, R. Verma, S. Sharma, S. Bhargava, A. Vahadane, and A. Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.
- [11] H. Irshad, L. Montaser-Kouhsari, G. Waltz, O. Bucur, J. Nowak, F. Dong, N. W. Knoblauch, and A. H. Beck, "Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd," in *Pacific Symposium on Biocomputing*, pp. 294–305, World Scientific, 2014.
- [12] P. Naylor, M. Laë, F. Reyat, and T. Walter, "Segmentation of nuclei in histopathology images by deep regression of the distance map," *IEEE TMI*, vol. 38, no. 2, pp. 448–459, 2018.
- [13] L. Hou, R. Gupta, J. S. Van Arnam, Y. Zhang, K. Sivalenka, D. Samaras, T. M. Kurc, and J. H. Saltz, "Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of 10 cancer types," *arXiv preprint arXiv:2002.07913*, 2020.
- [14] J. Gamper, N. A. Koohbanani, S. Graham, M. Jahanifar, S. A. Khurram, A. Azam, K. Hewitt, and N. Rajpoot, "Pannuke dataset extension, insights and baselines," *arXiv preprint arXiv:2003.10778*, 2020.
- [15] D. Liu, D. Zhang, Y. Song, F. Zhang, L. O'Donnell, H. Huang, M. Chen, and W. Cai, "Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting," in *CVPR*, pp. 4243–4252, 2020.
- [16] F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review," *IEEE reviews in biomedical engineering*, vol. 9, pp. 234–263, 2016.
- [17] T. Vicar, J. Balvan, J. Jaros, F. Jug, R. Kolar, M. Masarik, and J. Gumolec, "Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–25, 2019.
- [18] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [19] C. Ortiz de Solorzano, E. Garcia Rodriguez, A. Jones, D. Pinkel, J. W. Gray, D. Sudar, and S. J. Lockett, "Segmentation of confocal microscope images of cell nuclei in thick tissue sections," *Journal of Microscopy*, vol. 193, no. 3, pp. 212–226, 1999.
- [20] S. J. Keenan, J. Diamond, W. Glenn McCluggage, H. Bharucha, D. Thompson, P. H. Bartels, and P. W. Hamilton, "An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (cin)," *The Journal of pathology*, vol. 192, no. 3, pp. 351–362, 2000.
- [21] X. Zhou, F. Li, J. Yan, and S. T. Wong, "A novel cell segmentation method and cell phase identification using markov model," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 2, pp. 152–157, 2008.
- [22] S. W. Zucker, "Region growing: Childhood and adolescence," *Computer graphics and image processing*, vol. 5, no. 3, pp. 382–399, 1976.
- [23] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta informaticae*, vol. 41, no. 1, 2, pp. 187–228, 2000.
- [24] G. Lin, M. K. Chawla, K. Olson, J. F. Guzowski, C. A. Barnes, and B. Roysam, "Hierarchical, model-based merging of multiple fragments for improved three-dimensional segmentation of nuclei," *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, vol. 63, no. 1, pp. 20–33, 2005.
- [25] L. Yang, Z. Qiu, A. H. Greenaway, and W. Lu, "A new framework for particle detection in low-snr fluorescence live-cell images and its application for improved particle tracking," *IEEE transactions on biomedical engineering*, vol. 59, no. 7, pp. 2040–2050, 2012.

- [26] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.
- [27] R. Nock and F. Nielsen, "On weighting clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 8, pp. 1223–1235, 2006.
- [28] S. Kothari, Q. Chaudry, and M. D. Wang, "Automated cell counting and cluster segmentation using concavity detection and ellipse fitting techniques," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 795–798, IEEE, 2009.
- [29] C. Jung, C. Kim, S. W. Chae, and S. Oh, "Unsupervised segmentation of overlapped nuclei using bayesian classification," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2825–2832, 2010.
- [30] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.
- [31] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2009.
- [32] H. Chang, J. Han, A. Borowsky, L. Loss, J. W. Gray, P. T. Spellman, and B. Parvin, "Invariant delineation of nuclear architecture in glioblastoma multiforme for clinical and molecular association," *IEEE transactions on medical imaging*, vol. 32, no. 4, pp. 670–682, 2012.
- [33] X. Wu, M. Amrikachi, and S. K. Shah, "Embedding topic discovery in conditional random fields model for segmenting nuclei using multispectral data," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 6, pp. 1539–1549, 2012.
- [34] H. Kong, M. Gurcan, and K. Belkacem-Boussaid, "Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell splitting," *IEEE transactions on medical imaging*, vol. 30, no. 9, pp. 1661–1677, 2011.
- [35] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.
- [36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.
- [37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [38] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pp. 3–11, Springer, 2018.
- [39] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.
- [40] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, "Hematoxylin and eosin staining of tissue and cell sections," *Cold spring harbor protocols*, vol. 2008, no. 5, pp. pdb-prot4986, 2008.
- [41] S. K. Sadanandan, P. Ranefall, S. Le Guyader, and C. Wählby, "Automated training of deep convolutional neural networks for cell segmentation," *Scientific reports*, vol. 7, no. 1, pp. 1–7, 2017.
- [42] S. Lal, D. Das, K. Alabhy, A. Kanfode, A. Kumar, and J. Kini, "Nucleisegnet: Robust deep learning architecture for the nuclei segmentation of liver cancer histopathology images," *Computers in Biology and Medicine*, vol. 128, p. 104075, 2021.
- [43] J.-B. Lugagne, H. Lin, and M. J. Dunlop, "Delta: Automated cell segmentation, tracking, and lineage reconstruction using deep learning," *PLoS computational biology*, vol. 16, no. 4, p. e1007673, 2020.
- [44] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*, pp. 1989–1998, PMLR, 2018.
- [45] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2517–2526, 2019.
- [46] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4085–4095, 2020.
- [47] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 318–329, 2021.
- [48] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," *arXiv preprint arXiv:2112.04894*, 2021.
- [49] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," pp. 8801–8809, 2021.
- [50] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, D. N. M. Zhang, Shichuan, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, vol. 80, p. 102517, 2022.
- [51] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.
- [52] W. Cui, Y. Liu, Y. Li, M. Guo, Y. Li, X. Li, T. Wang, X. Zeng, and C. Ye, "Semi-supervised brain lesion segmentation with an adapted mean teacher model," in *International Conference on Information Processing in Medical Imaging*, pp. 554–565, Springer, 2019.
- [53] Z. Wang, Y. Wei, R. Feris, J. Xiong, W.-M. Hwu, T. S. Huang, and H. Shi, "Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 936–937, 2020.
- [54] K. Li, S. Wang, L. Yu, and P.-A. Heng, "Dual-teacher++: Exploiting intra-domain and inter-domain knowledge with reliable transfer for cardiac segmentation," *IEEE Transactions on Medical Imaging*, 2020.
- [55] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, "Adversarial learning for semi-supervised semantic segmentation," *arXiv preprint arXiv:1802.07934*, 2018.
- [56] X. Liu, F. Xing, C. Yang, G. El Fakhri, and J. Woo, "Adapting off-the-shelf source segmenter for target medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 549–559, Springer, 2021.
- [57] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.
- [58] M. Hu, T. Song, Y. Gu, X. Luo, J. Chen, Y. Chen, Y. Zhang, and S. Zhang, "Fully test-time adaptation for image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 251–260, Springer, 2021.
- [59] Z. Fu, J. Jiao, M. Suttie, and J. A. Noble, "Facial anatomical landmark detection using regularized transfer learning with application to fetal alcohol syndrome recognition," *IEEE Journal of Biomedical and Health Informatics*, 2021.
- [60] C. Chen, Q. Liu, Y. Jin, Q. Dou, and P.-A. Heng, "Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 225–235, Springer, 2021.
- [61] Y. Hou and L. Zheng, "Visualizing adapted knowledge in domain transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13824–13833, 2021.
- [62] F. Fleuret et al., "Uncertainty reduction for model adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9613–9623, 2021.
- [63] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.
- [64] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [65] S. Cui, S. Wang, J. Zhuo, L. Li, Q. Huang, and Q. Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3941–3950, 2020.
- [66] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [67] A. Obukhov, S. Georgoulis, D. Dai, and L. Van Gool, "Gated crf loss for weakly supervised semantic image segmentation," *arXiv preprint arXiv:1906.04651*, 2019.
- [68] C. Sutton, A. McCallum, et al., "An introduction to conditional random fields," *Foundations and Trends® in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.

- [69] M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, and C. Schroers, "Normalized cut loss for weakly-supervised cnn segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1818–1827, 2018.
- [70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.
- [71] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," IEEE Transactions on pattern analysis and machine intelligence, vol. 15, no. 9, pp. 850–863, 1993.
- [72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.
- [73] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in Proceedings of the European conference on computer vision (ECCV), pp. 289–305, 2018.
- [74] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," Advances in neural information processing systems, vol. 33, pp. 596–608, 2020.
- [75] X. Liu, C. Yoo, F. Xing, C.-C. J. Kuo, G. El Fakhri, J.-W. Kang, and J. Woo, "Unsupervised black-box model domain adaptation for brain tumor segmentation," Frontiers in Neuroscience, p. 341, 2022.