Source-Free Open-Set Domain Adaptation for Histopathological Images via Distilling Self-Supervised Vision Transformer

Guillaume Vray*, Devavrat Tomar*, Behzad Bozorgtabar*† and Jean-Philippe Thiran*†
*EPFL †CHUV
*{firstname}.{lastname}@epfl.ch

***Abstract*—There is a strong incentive to develop computational pathology models to i) ease the burden of tissue typology annotation from whole slide histological images; ii) transfer knowledge, e.g., tissue class separability from the withheld source domain to the distributionally shifted unlabeled target domain, and simultaneously iii) detect Open Set samples, i.e., unseen novel categories not present in the training source domain. This paper proposes a highly practical setting by addressing the abovementioned challenges in one fell swoop, i.e., source-free Open Set domain adaptation (SF-OSDA), which addresses the situation where a model pre-trained on the inaccessible source dataset can be adapted on the unlabeled target dataset containing Open Set samples. The central tenet of our proposed method is distilling knowledge from a self-supervised vision transformer trained in the target domain. We propose a novel style-based data augmentation used as hard positives for self-training a vision transformer in the target domain, yielding strongly contextualized embedding. Subsequently, semantically similar target images are clustered while the source model provides their corresponding weak pseudo-labels with unreliable confidence. Furthermore, we propose cluster relative maximum logit score (CRMLS) to rectify the confidence of the weak pseudo-labels and compute weighted class prototypes in the contextualized embedding space that are utilized for adapting the source model on the target domain. Our method significantly outperforms the previous methods, including open set detection, test-time adaptation, and SF-OSDA methods, setting the new state-of-the-art on three public histopathological datasets of colorectal cancer (CRC) assessment- Kather-16, Kather-19, and CRCTP. Our code is available at https://github.com/LTS5/Proto-SF-OSDA.**

***Index Terms*—Histopathological image analysis, Colorectal cancer assessment, data augmentation, Open-set domain adaptation, Vision transformer**

1. Introduction

Computational pathology has become a ripe ground for deep learning approaches as it has witnessed a rapid influx of myriad tasks, such as tissue phenotyping from whole slide images (WSIs). Nevertheless, even in routine clinical practice, curating huge-size WSIs with the heterogeneity of multiple tissues remains a daunting challenge.

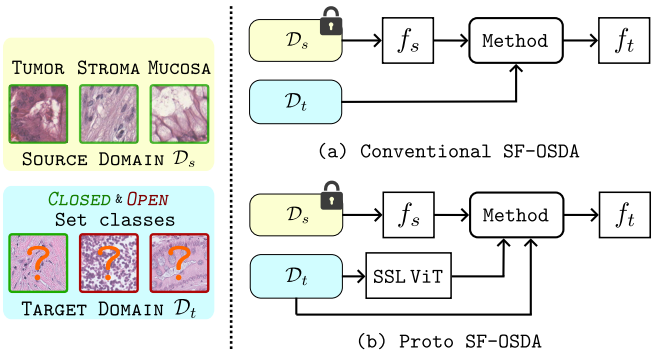

Figure 1: **Conceptual comparison of conventional SF-OSDA vs. Proto SF-OSDA**: (a) In conventional SF-OSDA, the source model f_s pre-trained on the inaccessible source domain $\mathcal{D}_s$ is adapted to the unlabeled target domain $\mathcal{D}_t$ comprising of Closed Set (known) and Open Set (unknown) classes; (b) Proto SF-OSDA can avail from a self-supervised vision transformer trained on the target domain, allowing access to a stronger contextualized embedding for the target data.

In this situation, transfer learning and domain adaptation techniques from open-source image datasets [1], [2] can mitigate or reduce costly annotation to be focused only on the labeling of the region of interest, i.e., relevant tissues. Albeit successful, access to a huge amount of labeled open-source data may be unattainable after deployment due to regulations on data privacy and computational limitations. This heralds a practical domain adaptation scenario, namely a source-free domain adaptation (SFDA) setting, where the task is to transfer knowledge from inaccessible source data to unlabeled target data using only a source-trained model. Several source-free domain adaptation methods [3], [4], [5] have been proposed en route to the goal based on test-time training/adaptation [6], [7]. However, while promising, these methods follow the closed-world assumption and do not account for the Open Set samples from target data that are not present in the training source set. The Open Set samples are unknown novel categories outside the set of initial categories annotated by the user, which could be **abundant but clinically irrelevant tissue categories** for the task at hand from histological slides or related to rare diseases and/or caused by acquisition artifacts. Thus, these

Open Set samples can be manually reviewed to avoid making the model susceptible to a potential misdiagnosis.

In addition to Open Set samples, the source-trained model may also encounter a shift in color appearance (chromatic shift) between the source and target domains of histopathology images, referred to as covariate shift due to different types of scanners or staining procedures. Nonetheless, current Open Set detection methods [8], [9], [10], [11] are often incapable of disentangling **semantically shifted Open Set samples (changes in the underlying structures, e.g., unknown tissue categories)** from the **covariate shifted Closed Set samples (known tissue categories with different chromatic distribution)** [12].

To this end, we propose Proto SF-OSDA, which focuses on a more practical Source-Free Open Set Domain Adaptation setting and mitigates the above challenges inherent in classifying histopathological tissue images in the presence of both covariate and semantic shifts. Contrary to previous SF-OSDA methods (Fig. 1 left), Proto SF-OSDA distills a self-supervised vision transformer to learn a stronger contextualized embedding for the target data (Fig. 1 right). Our major contributions are as follows:

- We propose automatic adversarial style augmentation (AdvStyle), which emulates covariate shifts resembling different chromatic staining shifts in histopathological images. The proposed data augmentation is utilized for self-supervised training of the vision transformer on the target data, yielding strongly contextualized embedding by pushing similar tissue features closer without using any labels.
- We propose Cluster Relative Maximum Logit Score (CRMLS) computed in the transformer's contextualized embedding space to refine the weak-pseudo labels with unreliable confidence from the source model. Consequently, using these refined confidence scores on the target images, we obtain weighted average class prototypes in the transformer's embedding space and generate reliable pseudo labels for adapting the source model.
- The proposed Proto SF-OSDA is evaluated on three histopathological datasets of CRC assessment. Experimental results demonstrate the superior performance of Proto SF-OSDA over the previous competing methods, including open set detection, test-time adaptation, and SF-OSDA methods under covariate and semantic shifts.

2. Related Work

2.1. Open Set Detection

In an Open Set setting, a trained model must be able to discriminate known (Closed Set) from unknown categories (Open Set) that have not yet been encountered. There are a few subcategories of Open Set recognition approaches. The first is model activation rectification strategies [13], [14], where Open Set examples can be detected using a rectification scheme for the activation patterns of the model related to Closed Set examples. Other approaches utilize GANs to generate images resembling Open Set samples [15] or train GAN-discriminator to distinguish closed from Open Set samples [16]. Seminal work [9] demonstrates that robust classifiers with high Closed Set accuracy also serve as better Open Set detectors. The authors establish Maximum Logit Score (MLS) as a baseline Open Set scoring metric for distinguishing known from unknown classes. Motivated by this, our method uses un-normalized logits for model adaptation and open-set recognition instead of the softmax output probabilities.

Other methods, such as Monte Carlo Dropout or Multi-head convolutional neural network (CNN) models [17], [18], measure uncertainty during inference to recognize Open Set images with high uncertainty. Another subcategory of methods evaluates the model on pretext tasks, e.g., predicting geometric [19] or color transformations [8] applied to test images during inference, with the assumption that the model will perform poorly on Open Set samples. Nonetheless, most of these methods require a specific source model training or access to Closed Set examples, limiting their applicability with off-the-shelf pre-trained models. Additionally, these methods assume similar image characteristics in the test domain, leading to failure under covariate shifts.

2.2. Source-Free Test-Time Domain Adaptation

Test-time domain adaptation, known as source-free domain adaptation (SFDA) [4], [20], [21], aims to improve model robustness against distribution shifts during inference without access to labeled source domain images. While current pseudo-label-based SFDA methods [4], [21], [22], [23], [24] work well in closed-world settings, they fail significantly in open-world scenarios and require filtering of Open Set examples during inference.

Moreover, another subcategory of SFDA methods, e.g., [25], [26], [27], solve particular auxiliary tasks (e.g., self-supervised rotation prediction) to adapt the model under common distributional shifts. Some other methods adapt batch normalization (BN) statistics [6] or proposed test-time augmentation to simulate augmentations resembling the saved BN statistics [7] to mitigate errors from Open Set examples. However, as shown in our experiments, solely bridging the target-source feature statistics gap may not be sufficient for improving Open Set recognition capabilities.

2.3. Source-Free Open Set Domain Adaptation

Existing SF-OSDA methods use self-supervised pseudo-labeling [4], uncertainty quantification in the source model prediction [28], or proposed specialized source training strategy [29] to train an inheritable model capable of adapting to the target domain with novel categories. In [4], the authors introduced a source-free domain adaptation method that involves a self-supervised pseudo-labeling scheme by clustering known and unknown categories using K-Means clustering. Subsequently, the model is adapted exclusively

on examples from the known categories in the target domain. Recently, a balanced progressive graph learning framework in introduced in [30] that decomposes the target hypothesis space into shared and unknown subspaces, employing progressive pseudo-labeling for hypothesis adaptation. In [31], inter-class relationships are modeled using the weights of the classifier layer of the source model, and this information is combined with contrastive learning to pseudo-label target domain images for adaptation.

In contrast to the previous works, we employ a self-supervised vision transformer to separately learn a contextualized embedding of target images and refine their corresponding weak-pseudo labels obtained from the source model. Our method does not rely on special source training strategies and can be applied to any off-the-shelf pre-trained source model.

3. Materials and Methods

Adapting a source model pre-trained on the inaccessible source data to the unlabeled target data of histopathological images under simultaneous *covariate* and *semantic* shifts is extremely challenging as the uncertainty of the source model's predictions on the target domain's images may come from either or both types of distributional shifts.

Overcoming this challenge, we propose distilling knowledge from a separate vision transformer (ViT) [32] based feature extractor trained on the unlabeled target domain's images in a self-supervised manner. This is motivated firstly because ViT-based models demonstrate better robustness against, e.g., texture bias than CNN-based models [33]. Secondly, self-supervised vision transformers, e.g., [34], yield a strong contextualized embedding space, which suggests the clustering approach in its feature space would function well. Nonetheless, recent self-supervised ViT-based methods, in particular, DINO-ViT [34], are trained by learning representational invariances to different augmented views of the natural RGB images, which may be sub-optimal to represent diverse color appearances simulating substantial staining variations of tissue samples. Motivated by the data augmentation policies used in [8], we propose using *automatic adversarial style augmentation* by learning the magnitudes of the color transformations for self-training the ViT on the target domain images. However, unlike [8], which utilizes an auxiliary network for test-time image transform prediction, we automatically learn image style-based adversarial augmentation policies to self-train ViT. The generated style-based augmentation policies are treated as *hard positives*, making distinguishing image pairs in the ViT feature space difficult. Consequently, the self-trained ViT provides contextualized embeddings of the target domain's images, while the original pre-trained source model provides their corresponding weak pseudo-labels. We then propose to refine the confidence of weak pseudo-labels by exploiting the semantic smoothness hypothesis of ViT's contextualized embedding space via CRMLS score and computing weighted class prototypes in this contextualized embedding space. Eventually, the ViT-guided class prototypes are used for adapting the source model to the target domain.

In summary, Proto SF-OSDA has two major components: (1) Self-supervised training of vision transformer via proposed automatic adversarial style augmentations to generate contextualized embedding space on the unlabeled target domain, and (2) Computing class prototypes using proposed CRMLS in the ViT embedding space, followed by adapting the source model on the obtained pseudo labels.

Problem formulation. Let $f_s : X \rightarrow Y$ denote the tissue-type classifier (source model) that is trained on the inaccessible source domain $\mathcal{D}_s$ of tissue images belonging to C known classes, where $X \in \mathbb{R}^{H \times W \times 3}$ denotes an RGB image with height H and width W, and $Y \in \mathbb{R}^C$ denotes the corresponding logits vector. Also, let $\mathcal{D}_t$ denote the unlabeled target domain images belonging to the C known and $\bar{C}$ unknown classes not seen during source model training, i.e., $C \cup \bar{C}$. We aim to adapt pre-trained source model f_s on $\mathcal{D}_t$ to obtain adapted model f_t so that it can correctly classify target domain images into C known classes and an Open Set category of samples containing $\bar{C}$ unknown classes. The Open Set target domain images are detected by comparing their maximum logit scores predicted by f_t against a threshold, while Closed Set images are classified into C known classes.

In Sec. 3.1, we describe our proposed adversarial style augmentation for self-supervised training of the vision transformer to obtain contextualized target embeddings. Sec. 3.2 presents our proposed CRMLS for refining the confidence of the source model on the target domain with Open Set examples, while Sec. 3.3 describes utilizing CRMLS scores to obtain Open Set aware class prototypes in the contextualized target domain embedding and final adaptation of the source model.

3.1. Self-Supervised ViT Training via Automatic Adversarial Style Augmentation

Let $\mathcal{F} : X \rightarrow Z$ denote the ViT feature extractor that maps the target images $X \in \mathbb{R}^{H \times W \times 3}$ to its contextualized embedding space $Z \in \mathbb{R}^d$. We adopt DINO [34] for self-supervised training of $\mathcal{F}$ on the target domain images $\mathcal{D}_t$ as it is a powerful nearest neighbor classifier, and thus better clusters semantically similar images in its embedding space. DINO uses an identical architecture of teacher-student networks, where the teacher network is slowly updated as a moving average by the student network during training. The soft-maxed logits predicted by the teacher network on randomly augmented global crops of the image views are matched by the student network on another set of augmented image views with global and local image crops. However, we replace the DINO default augmentations with our automatic adversarial style augmentation (AdvStyle) for learning contextualized embedding of the target images.

Automatic adversarial style augmentation (AdvStyle). Using the data augmentation policies from [8], let $\mathbb{O}$ represent the set of color transformations operations $\mathcal{O} : X \rightarrow X$.

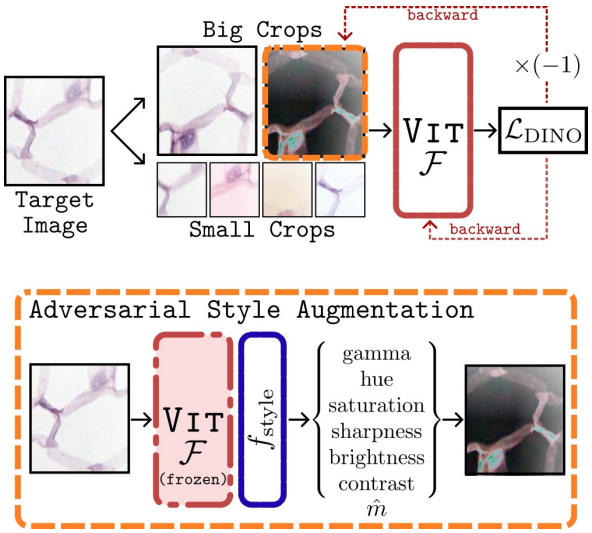

Figure 2: **Self-training vision transformer via proposed adversarial style augmentations.** The. style augmentation module f_{style} is trained to learn magnitudes $\hat{m}$ for adversarial augmentations by maximizing $\mathcal{L}_{\text{DINO}}$ while transformer encoder $\mathcal{F}$ is updated to minimize $\mathcal{L}_{\text{DINO}}$ on the target domain.

We define $\mathbb{O}$ as the set of six color transformations, including $\{$Gamma, Hue, Saturation, Sharpness, Brightness, Contrast$\}$. Each color transformation in $\mathbb{O}$ is applied to a given tissue image parametrized by its magnitude $\hat{m} \in [0, 1]$, which determines the strength of the color transformation. We utilize a style augmentation module consisting of a shallow multilayer perception (MLP) f_{style} built on top of the frozen ViT teacher network that sequentially applies differentiable image transformation operations $\mathcal{T}(x, \hat{m})$ to input image x using their predicted magnitudes $\hat{m}$ as follows:

$$\hat{x} = \mathcal{T}(x, \hat{m}) = \mathcal{O}_6(\mathcal{O}_5...(\mathcal{O}_1(x, \hat{m}_1), ..., \hat{m}_5, \hat{m}_6) \quad (1)$$

where $\mathcal{O}_j(x, \hat{m}_j)$ applies j^{th} color transform with learnable magnitude $\hat{m}_j$ on a given target domain image x to generate augmented image $\hat{x}$.

As shown in Fig. 2, we self-train ViT encoder $\mathcal{F}$ on $\mathcal{D}_t$ using the same self-supervised DINO loss $\mathcal{L}_{\text{DINO}}$ (Eq. 2) as [34] but with our AdvStyle instead of their default augmentations. f_{style} is trained in an adversarial manner w.r.t. ViT encoder to generate hard style-based augmentation policies that make distinguishing augmented image pairs in the feature space difficult. In this adversarial setup, f_{style} is trained to maximize $\mathcal{L}_{\text{DINO}}$ (Eq. 2), while the ViT model is trained to minimize $\mathcal{L}_{\text{DINO}}$. Note that we only learn to adversarially augment one of the two global image crops while randomly augmenting the global and local crops using the default color jittering as shown in Fig. 2. $\mathcal{L}_{\text{DINO}}$ is defined as follows:

$$\mathcal{L}_{\text{DINO}}(x) = \sum_{i=1}^{2} \Big(\sum_{\substack{j=1 \\ j \neq i}}^{2} H(G_t(x_g^i), G_s(x_g^j)) + \sum_{k=1}^{4} H(G_t(x_g^i), G_s(x_l^k)) \Big) \quad (2)$$

where,

$$x_g^1 = \text{Crop}_g(\text{aug}(x)), \quad x_g^2 = \mathcal{T}(x, f_{\text{style}}(\mathcal{F}_t(x)), \quad x_l^k = \text{Crop}_l(\text{aug}(x))$$

Here, G_t and G_s denote the teacher, and student networks, each composed of a feature extractor $\mathcal{F}$ and a projection head $g : Z \rightarrow U$ applied on $\mathcal{F}$'s output $Z, U \in \mathbb{R}^{d'}$. $\text{Crop}_g(.)$ and $\text{Crop}_l(.)$ denote random global and local crops, while $\text{aug}(.)$ denotes random color jittering transformation applied to the target domain's images with a probability of 0.5. $H(p, q) = -\sum p \log q$ denotes the cross-entropy loss.

The parameters of the student network θ_{G_s} and that of adversarial augmentation module $\theta_{f_{\text{style}}}$ are updated to optimize the following mini-max objective:

$$\min_{\theta_{G_s}} \max_{\theta_{f_{\text{style}}}} \mathbb{E}_{x \sim \mathcal{D}_t} \left[\mathcal{L}_{\text{DINO}}(x) \right] \quad (3)$$

The teacher network's parameters θ_{G_t} are updated by that of the student network θ_{G_s} with momentum $\nu = 0.996$ after every gradient step as follows:

$$\theta_{G_t} = \nu \cdot \theta_{G_t} + (1 - \nu) \cdot \theta_{G_t} \quad (4)$$

3.2. Cluster Relative Maximum Logit Score

We utilize the contextualized embedding space of self-supervised trained ViT ($\mathcal{F}$) to correct the confidence of the source model f_s on the target domain of tissue mages. As the representation of the target domain's images in the $\mathcal{F}$'s embedding space is semantically smooth (images belonging to the same tissue category are mapped to nearby points in embedding space and have similar confidence), we rectify the source model's confidence on the target domain's images by assigning similar confidence scores to the images that are neighbors in the $\mathcal{F}$'s embedding space. To achieve this, we first group the contextualized embeddings of the target domain images obtained by $\mathcal{F}$ into a large number of groups using a simple K-Means clustering, followed by assigning the same *cluster-wise mean logit vector* (given by f_s) of size C (number of known classes) to all the images belonging to the same cluster. The cluster relative confidence of the source model for a target domain's image is then given by the maximum logit value of its corresponding *cluster-wise mean logit vector*. Note that we use the unnormalized maximum raw logit output of f_s to measure its confidence in a target domain image instead of the maximum softmax probability as the former has been shown more effective in detecting Open Set examples [9]. In addition, we re-calibrate the batch normalization [35] statistics of the source model f_s with that of the target domain if f_s is a standard CNN. Re-calibrating

batch normalization statistics of f_s can effectively reduce the feature distribution shifts encountered in CNNs [6].

Clustering of the target domain's contextualized embeddings. As shown in Fig. 3, we use K-Means clustering to group the contextualized embeddings of the target domain images $\mathcal{D}_t$ obtained from $\mathcal{F}$ into K disjoint clusters represented by $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$, where K is chosen arbitrarily large, e.g., 16 (see ablation in Fig. 8). Each cluster member is then assigned the same confidence score, namely cluster relative maximum logit score (CRMLS).

Computing CRMLS. The CRMLS scores of all target domain's images belonging to the cluster $\mathcal{D}_i$ are computed as:

$$\Phi_{\text{CRMLS}}(\mathcal{D}_i) \leftarrow \max \mathbb{E}_{x \sim \mathcal{D}_i}\{f_s(x)\} \tag{5}$$

where $\max$ is taken over C known classes, and $\mathbb{E}$ is the expectation operation. CRMLS helps to correct f_s's confidence by assigning the expected confidence score to the target domain images belonging to the same cluster in the transformer's contextualized embedding space. The target domain images in a cluster with homogeneous and confident predictions by f_s receive high scores, while images of heterogeneous or unconfident clusters receive low scores.

3.3. Source-Free Adaptation via Class Prototypes

To obtain the pseudo-labels for self-training the source model f_s on the target domain images, we first compute the class prototypes $P = [p_1, p_2, \ldots p_C]$ of C known classes in the transformer's contextualized embedding space using CRMLS scores. Let $X^j = \{\forall x | x \in \mathcal{D}_t \wedge \arg\max f_s(x) = j\}$ denote all target domain images that are classified by f_s as class j. Also, $\forall x \in X^j$, let $\Phi_{\text{CRMLS}}(x)$ denote its confidence score and $S_{\text{cluster}}(x) = |\mathcal{D}_{i \ni x \in \mathcal{D}_i}|$ denote the size of the cluster containing x. We compute the prototype of class j as:

$$p_j = \frac{\sum_{x \in X^j} \hat{w}(x) \cdot \mathcal{F}(x)}{\sum_{x \in X^j} \hat{w}(x)} \tag{6}$$

where,

$$\hat{w}(x) = w(x) - \min_{x \in \mathcal{D}_t} w(x), \qquad w(x) = \frac{\Phi_{\text{CRMLS}}(x)}{S_{\text{cluster}}(x)}$$

Due to instability caused by K-Means clustering initialization, we run N_{mc} Monte-Carlo simulations of K-Means clustering (see ablation in Fig. 8) to compute $\Phi_{\text{CRCLS}}(x)$ and weights $w(x)$ of the target domain images $x \in \mathcal{D}_t$ to get better estimates. $\hat{w}(x)$ ensures the weights are positive and dividing $\Phi_{\text{CRCLS}}(x)$ by the corresponding cluster size $S_{\text{cluster}}(x)$ ensures the prototypes are not biased towards large clusters.

Let $P_{\text{norm}} = [\hat{p}_1, \hat{p}_2, \ldots \hat{p}_C]$ denote unit norm class prototypes such that $\hat{p}_j = p_j / \|p_j\|$. We compute the unnormalized pseudo-logits $\bar{y}$ for the target image x as follows:

$$\bar{y}(x) = P_{\text{norm}}^T \frac{\mathcal{F}(x)}{\|\mathcal{F}(x)\|_2} / \tau \tag{7}$$

where τ is a temperature hyperparamter. The source model f_s is adapted on the target domain images $\mathcal{D}_t$ using the pseudo-logits of Eq. 7 to obtain f_t^* by minimizing the following mean squared error (MSE) loss:

$$f_t^* = \arg\min_f \mathbb{E}_{x \sim \mathcal{D}_t} \|f(x) - \bar{y}(x)\|_2^2 \tag{8}$$

We opt for MSE loss instead of the KL-divergence loss as the former allows better Open Set detection based on *Maximum Logit Score* (see ablation in Table 3).

4. Experiments and Results

4.1. Datasets

We evaluate Proto SF-OSDA on CRC tissue phenotyping using Hematoxylin and Eosin (H&E) stained tissue sections extracted from WSIs of colorectal biopsies. In particular, we utilize three publicly available CRC tissue characterization datasets digitized at a magnification of 20×: Kather-16 [36], Kather-19 [1], and CRCTP [2]. The Kather-16 dataset comprises 5,000 patches (150×150 pixels, 0.495 µm/pixel) representing eight tissue classes, with a balanced distribution of 625 patches per class. Kather-19 consists of 100,000 tissue patches (224×224 pixels, 0.5 µm/pixel) divided almost evenly among nine classes. CRCTP contains 196,000 image patches (150×150, 0.495 µm/pixel) categorized into seven tissue phenotypes. For the experiments, we apply a random stratified split to divide all datasets into training (70%), validation (15%), and test (15%) sets. Due to discrepancies in class definitions, following consultation with expert pathologists, we follow the same harmonization approach in [37], resulting in a set of 7 common classes: tumor epithelium (TUM), stroma (STR), lymphocytes (LYM), normal colon mucosa (NORM), complex stroma (c-STR), debris (DEB), and background (BACK). To make correspondence between datasets, following [37], we merge stroma and smooth muscle (MUS) classes as stroma (STR); and debris and mucus (MUC) as debris (DEB).

Closed and Open Set splits. We adopt the experimental setup outlined in [8] to define Closed and Open Set splits for each dataset. The three splits depicted in Fig. 4 follow the same principles described in [8], with minor modifications. Specifically, Split 1 emulates a scenario where a practitioner labels only clinically relevant tissue regions (TUM, STR, LYM, and NORM) used for the critical task of, e.g., quantifying tumor-stroma ratio [38], [39], leaving uninformative regions (c-STR, DEB, ADI, BACK) unlabeled. Notably, in contrast to [8], c-STR is considered an Open Set category because it is not present in Kather-19 and CRCTP datasets [37]. Split 2 and Split 3 complement the analysis by focusing on the classification of tumoral regions (TUM and STR) while excluding healthy tissues (LYM, MUC) and uninformative but abundant samples (c-STR, DEB, ADI, BACK). Additionally, Split 3 includes lymphocyte images (LYM) in the Closed Set of Split 2 to introduce more challenging Closed Set classification scenarios. Table

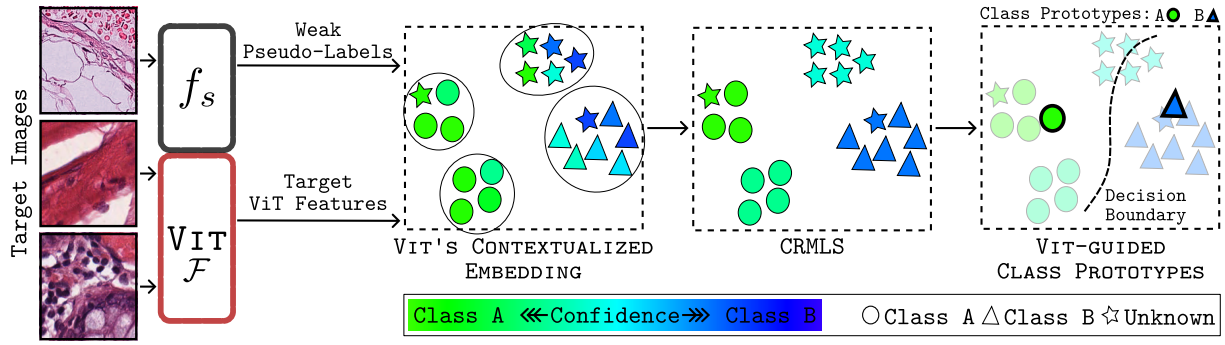

Figure 3: **Transformer guided class prototypes.** Given self-supervised ViT $\mathcal{F}$ and the source model f_s, we first obtain the target images' representations in $\mathcal{F}$'s contextualized embedding space which are weakly labeled by f_s and grouped into K clusters using K-Means clustering. Using CRMLS scores (Sec. 3.2), we then refine the confidence of weak pseudo labels and compute weighted class prototypes of known classes in $\mathcal{F}$'s embedding space.

TABLE 1: **Summary of datasets repartitions.** For all the datasets used in our experiments, we report the class ratio and number of image patches in train, validation, and test sets of every Closed and Open Set split – Splits 1, 2, and 3.
* The validation subset of Open Set is not used in our experiments.

		Kather-16				Kather-19				CRCTP			
			Train	Validation	Test		Train	Validation	Test		Train	Validation	Test
		Ratio	70%	15%	15%	Ratio	70%	15%	15%	Ratio	70%	15%	15%
Split 1	Closed Set	50.0%	1,756	372	372	58.6%	41,039	8,790	8,790	92.9%	127,400	27,300	27,300
	Open Set	50.0%	1,756	372*	372	41.4%	28,971	6,205*	6,205	7.1%	9,800	2,100*	2,100
Split 2	Closed Set	25.0%	878	186	186	38.3%	26,813	5,743	5,743	71.4%	98,000	21,000	21,000
	Open Set	75.0%	2,634	558*	558	61.7%	43,197	9,252*	9,252	28.6%	39,200	8,400*	8,400
Split 3	Closed Set	37.5%	1,317	279	279	49.9%	34,904	7,476	7,476	82.1%	112,700	24,150	24,150
	Open Set	62.5%	2,195	465*	465	50.1%	35,106	7,519*	7,519	17.9%	24,500	5,250*	5,250

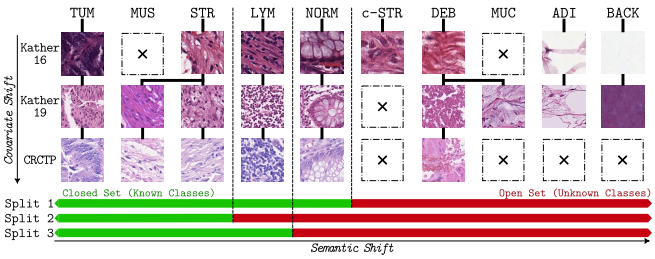

Figure 4: **Summary of dataset partitions** used for Closed (green) and Open (red) Set defined on the three CRC datasets for tissue-type classification.

1 reports each dataset repartition and size, along with the corresponding Closed and Open Set class ratios. The Open Set class ratio ranges from 7.1% to 75%, allowing us to cover various SF-OSDA scenarios.

4.2. Evaluation Protocol

We evaluate all baselines and our proposed Proto SF-OSDA following the same protocol. Firstly, we train the source model f_s on the training Closed Set and validate it on the validation Closed Set of the source domain for selecting the best-performing source model. Next, we adapt $f_s \rightarrow f_t$ on the target domain's training subset, including Closed and Open Sets. Finally, we evaluate the performance of f_t on the test subset of the target domain. We use class average Closed Set accuracy (ACC) to assess domain adaptation on the Closed Set and area under the ROC curve (AUC) to access Open Set detection. We perform six adaptation experiments, including pairwise adaptation among Kather-16, Kather-19, and CRCTP datasets on every split. We report results averaged over ten random seeds of source model training.

4.3. Baselines

We compare our method against several seminal and state-of-the-art (SOTA) methods for Open Set detection (OSR), source-free test-time domain adaptation (TTA), and SF-OSDA. Regarding OSR baselines, we compare our method against CE+ [9] (maximizing the Closed Set accuracy of the source model) and MC-Dropout [18]. In addition, we adopt SOTA OSR baselines, including T3PO [8] and CSI [11] as representative of augmentation-based approaches. The former uses test-time image transformation prediction, while the latter leverages contrastive learning to contrast the sample with distributionally-shifted augmentations of itself. For TTA baselines, we opt for parameter-free techniques, including test-time batch adaptation techniques, BN [6], and test-time augmentation method, namely OptTTA [7]. Finally,

TABLE 2: **Performance comparison against state-of-the-art methods** on Closed Set domain adaptation (ACC↑ in %) and Open Set detection (AUC↑ in %) for pairwise model adaptation on Kather-16, Kather-19, and CRCTP datasets. Results are averaged over 10 seeds. [‡] $p < 0.001$, [†] $0.001 \leq p < 0.1$, [n] $p \geq 0.1$; paired t-test with respect to top results.

		Open Set Detection Methods				**TTA Methods**		**SF-OSDA Methods**		
Split	Metric	CE+	MC-Dropout	T3PO	CSI	BN	OptTTA	SHOT	U-SFAN	Proto SF-OSDA
Kather-19 → Kather-16 / Kather-16 → Kather-19										
1	ACC	75.3/88.6	76.4/88.2	89.2/92.9	79.3/89.2	89.0/86.3	82.9/79.7	88.5/89.6	89.1/87.5	**94.3**‡**/98.0**‡
	AUC	85.4/78.0	80.5/81.1	88.4/80.7	66.4/74.1	86.3/69.8	71.4/61.5	80.5/51.4	85.3/61.1	**95.7**‡**/99.3**‡
2	ACC	97.9/75.7	98.2/77.0	98.8/89.1	95.2/92.1	98.0/88.0	96.4/79.1	97.6/90.1	97.7/88.1	**99.3**†**/98.7**‡
	AUC	86.1/66.7	86.0/67.7	85.6/78.8	72.2/61.9	88.3/71.9	78.3/68.9	83.8/65.8	87.1/65.4	**96.0**‡**/92.2**‡
3	ACC	84.5/73.4	84.9/81.4	88.9/91.0	70.1/94.0	88.3/87.1	85.8/85.0	85.7/86.4	88.0/84.9	**93.6**‡**/99.0**†
	AUC	83.0/65.1	82.2/64.5	85.6/78.6	66.1/55.8	87.6/80.7	76.1/80.2	73.9/64.3	85.0/65.7	**95.2**‡**/98.8**‡
CRCTP → Kather-16 / Kather-16 → CRCTP										
1	ACC	71.0/67.3	72.1/68.4	72.0/76.5	66.4/61.9	83.3/74.1	79.3/74.3	87.8/76.0	84.4/74.1	**95.2**‡**/80.2**‡
	AUC	69.6/67.9	64.9/70.6	82.8/60.8	80.9/65.0	76.5/61.7	72.8/60.6	78.7/61.7	76.2/61.6	**94.4**‡**/80.4**‡
2	ACC	97.1/85.2	96.8/84.2	81.4/92.8	95.2/84.3	98.1/94.2	95.1/89.3	97.5/95.0	98.0/94.4	**99.2**†**/95.8**‡
	AUC	76.9/70.2	75.0/69.0	72.6/70.7	72.2/61.4	75.6/78.3	61.2/75.2	72.8/76.0	75.2/77.9	**92.3**‡**/89.8**‡
3	ACC	89.0/66.9	86.3/70.0	83.7/81.9	70.1/59.2	94.5/81.2	91.6/80.2	95.3/82.7	95.0/82.0	**95.9**†**/84.3**†
	AUC	73.1/69.4	66.4/68.4	82.1/70.3	66.1/64.3	72.6/72.0	73.0/69.8	71.9/71.8	72.5/72.1	**89.2**†**/87.6**‡
CRCTP → Kather-19 / Kather-19 → CRCTP										
1	ACC	73.9/52.5	75.4/52.5	75.6/53.1	72.2/60.7	82.1/73.0	68.4/62.9	84.6/74.7	89.1/73.7	**96.8**‡**/80.5**‡
	AUC	72.0/63.9	76.7/63.6	85.9/61.8	72.5/64.4	75.7/61.2	60.0/64.9	66.4/56.7	56.9/52.6	**98.3**‡**/78.4**‡
2	ACC	95.2/88.6	95.8/88.5	94.9/87.5	93.9/93.1	92.1/93.6	85.9/92.8	92.2/94.1	90.6/94.4	**99.4**‡**/95.8**‡
	AUC	73.6/70.6	73.4/72.0	74.0/68.2	71.3/77.1	77.2/82.3	67.0/76.8	68.6/78.3	58.9/71.4	**95.1**‡**/90.2**‡
3	ACC	94.4/57.7	94.3/57.2	94.9/59.7	94.9/62.1	94.3/78.2	88.1/66.5	93.9/80.9	92.2/82.9	**99.1**‡**/85.1**‡
	AUC	73.5/79.2	75.9/74.3	81.9/77.9	85.3/78.4	78.0/74.6	71.4/68.5	68.7/69.5	61.1/62.8	**97.9**‡**/84.3**‡

we compare with SOTA SF-OSDA baselines, including SHOT [4], and U-SFAN [28]. For CE+, MC-Dropout, and T3PO, we use the released codes by [8], which are slightly adapted to our setting, while for other methods, we use their implementation with recommended parameters for a fair comparison.

4.4. Implementation Details

We adopt the ViT/B-16 encoder as $\mathcal{F}$ starting from a self-supervised DINO [34] initialization and MobileNet-V2 [41] for the source-trained model f_s. We train the source model f_s for 100 epochs on Kather-16, 20 epochs on Kather-19, and CRCTP, respectively, for all splits using the training strategy mentioned in CE+ [9]. For self-training the vision transformer on the target domain, we use AdamW [42] optimizer with a learning rate of 2.5e-4 for 40, 5, and 10 epochs on Kather-16, Kather-19, and CRCTP datasets, respectively. We compute the class prototypes using our proposed CRMLS with $K = 16$ clusters, $N_{\text{mc}} = 32$ Monte-Carlo simulations (see ablation in Fig. 8), and temperature $\tau = 0.07$. Finally, for the model adaptation step, we self-train f_s on the target domain with the obtained log pseudo-labels for five epochs on Kather-16 and two epochs otherwise, using Adam optimizer [43] with a learning rate of 1e-3. The experiments were carried out using PyTorch 1.13 on an NVIDIA GeForce GTX 1080 Ti GPU with 12GB of memory.

4.5. Comparisons with State-of-the-Art Methods

Table 2 compares Proto SF-OSDA against SOTA methods for Closed Set domain adaptation scenario via ACC metric and Open Set recognition capability through AUC metric on three splits for adapting the source model from Kather-19 to Kather-16 and CRCTP to Kather-16, and vice-versa. Proto SF-OSDA consistently achieves the highest AUC and ACC scores for all three CRC tissue datasets across both adaptation scenarios. In particular, our Proto SF-OSDA notably achieves significant performance gain on Open Set recognition (+18.1% AUC in Split 3) and Closed Set domain adaptation scenario (+8.4% ACC in Split 1) for the Kather-16 → Kather-19 model adaptation.

Several conclusions can be drawn based on the experimental comparison to other methods. First, one can observe that Open Set detection methods lag behind Proto SF-OSDA when exposed to covariate shifts due to variations in tissue's visual appearances. Second, TTA methods do not perform well under semantic shifts as these methods assume a closed-world setup. Third, similar poor performance trends are observed for recent SF-OSDA methods, including U-SFAN and SHOT. Both SHOT and U-SFAN detect Open Set samples by analyzing the source model's features which may not be very informative for Open Set detection to handle severe chromatic shifts of tissue patches. Even worse, these methods require much more target domain data for the

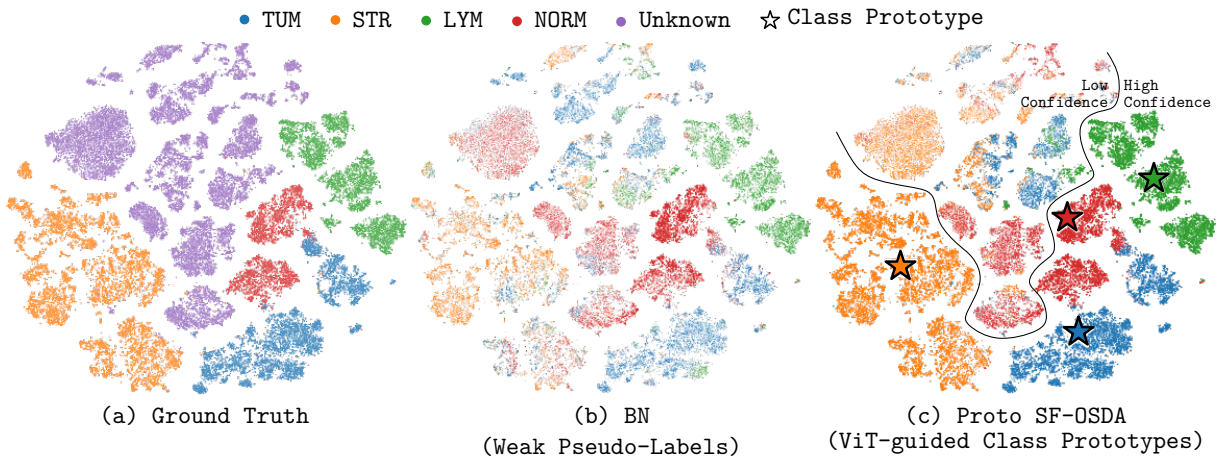

Figure 5: **The t-SNE visualization [40] comparison** of self-supervised trained transformer encoder $\mathcal{F}$ of feature embeddings on target domain images (Kather-19, Split 1), which are color-coded to distinguish different tissue types: **(a)** Ground Truth; **(b)** BatchNorm re-calibrated source model f_s trained on the Kather-16; and **(c)** Proto SF-OSDA's feature embedding. The weighted average class prototypes are obtained in the transformer's embedding space. *Labels' confidence* $\propto$ *Color Density*.

model adaptation. We argue that the superior performance of Proto SF-OSDA is a direct effect induced by knowledge distillation through a separate self-supervised trained vision transformer in the distributionally shifted unlabeled target domain. The quantitative results are supported by the t-SNE [40] visualization results (Fig. 5) of self-supervised trained ViT encoder $\mathcal{F}$ of feature embeddings on target domain images (Kather-19, Split 1). Our transformer-guided class prototypes successfully refine the weak pseudo-labels (given by f_s) of the target domain images in the ViT encoder's feature space with reliable confidence in the presence of Open Set samples and covariate shift caused by changes in tissue color appearance.

Tissue image segmentation. To demonstrate the capability of Proto SF-OSDA for segmentation on large-resolution tissue-level images, we adapted the patch-level classifier for the Kather-19→Kather-16 model adaptation. In Fig. 6, we show our method's pixel-level segmentation results on a large tissue image (5000×5000 pixels) alongside corresponding segmentation maps generated by competing methods, i.e., CE+ and BN. For each method, we present the segmentation maps computed using the predicted logit score for tumor (TUM) and MLS score for Open Set and Closed Set tissue classes. In the absence of ground truth information for the presented image, our method is seen to better segregate Closed set tissue classes (TUM, STR, LYM, MUC) from Open Set classes (c-STR, DEB, ADI, and BACK) (Fig. 6, bottom). In addition, Proto SF-OSDA produces a less noisy segmentation map, yielding better delineation of tumor regions than other methods (Fig. 6, top).

4.6. Ablation Study

Style-based data augmentation. We evaluated the effect of our proposed style-based augmentation (AdvStyle) (Sec. 3.1) for self-supervised training of ViT against the standard

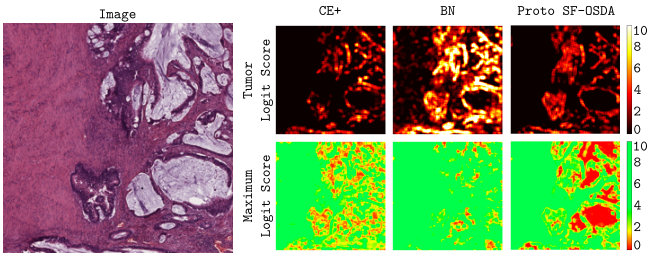

Figure 6: **Tissue image segmentation results** for the Kather-19→Kather-16 model adaptation performed by CE+, BN, and Proto SF-OSDA on unseen tissue image: (top) predicted logit score for the tumor (TUM) class; (bottom) predicted MLS score, where green and red regions correspond to Closed set classes (TUM, STR, LYM, MUC), and Open Set classes (c-STR, DEB, ADI and BACK), respectively.

augmentation policies by comparing the quality of final pseudo-labels. Fig. 7(a) shows that our AdvStyle policies perform much better against augmentation policies used in DINO [34], i.e., random color-jittering, and T3PO [8] on every Closed/Open Set split on the Kather-16→Kather-19 adaptation.

Weighted average class prototypes, sensitivity to K-Means initialization/clusters. In Fig. 7(b), we show the effectiveness of our CRMLS scores for computing weighted average class prototypes of Eq. 6 in the contextualized embedding space of ViT by comparing it against uniform scoring (US), and MLS [9]. We observe that CRMLS produces superior-quality pseudo-labels. Moreover, since Proto SF-OSDA is based on K-Means clustering, which is highly sensitive to initialization and K, we perform a sensitivity test on the number of Monte-Carlo runs N_{mc} of K-Means and the number of clusters K in Fig. 8. We observe that Proto SF-OSDA is insensitive to these parameters after

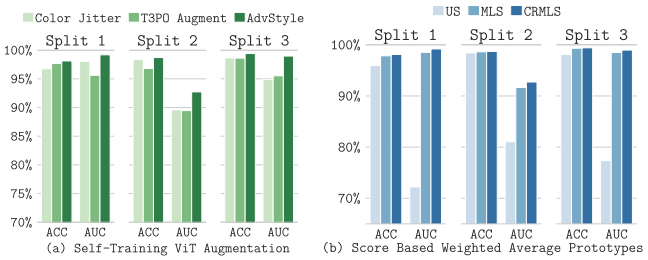

Figure 7: **Ablation studies on the Kather-16→Kather-19 adaptation**. **(a)** Style-based augmentation types used for self-training of ViT on the target domain: Random Color-Jittering, T3PO's [8], and our AdvStyle; **(b)** Scores for computing weighted average class prototypes: uniform score (US), maximum logit score (MLS), and CRMLS.

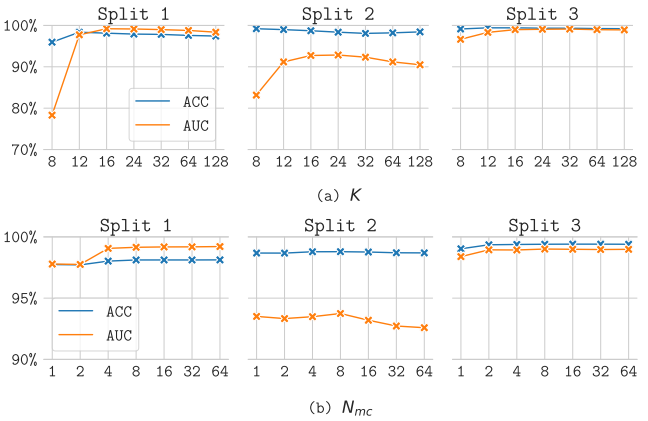

Figure 8: **Ablations** on the number of clusters K for K-means clustering and the number of Monte-Carlo runs N_{mc} for Kather-16 → Kather-19 adaptation.

a certain threshold, e.g., $K = 12$, $N_{mc} = 4$.

Distillation loss for source model adaptation. In this ablation study, we investigate different distillation loss functions for adapting the source model f_s on the training set of the target domain using the corresponding prototypical pseudo-logits of Eq. 8. In Table 3, we report the performance of the model f_t* on the test set of the target domain adapted using traditional cross-entropy loss (CE), SHOT [4]'s InfoMax with CE, l_1 and the suggested l_2 loss for self-distillation. All distillation losses give similar ACC; however, only l_1 and l_2 provide high AUC scores. This is attributed to the fact that l_1 and l_2 losses also penalize the magnitude of the logits given by the target-adapted model, thus better preserving the MLS-based confidence ranking of the target domain images.

5. Practicality of Proto SF-OSDA

This section discusses the practicality of Proto SF-OSDA for clinical feasibility regarding source model architecture sensitivity, amount of target data, and runtime. As shown in Fig. 1 (a), SF-OSDA methods require two inputs: a pre-trained source model f_s and unlabeled data from the target domain. We argue the source model architecture should be unknown a priori; therefore, SF-OSDA methods must be insensitive to model architecture. The target data's quantity is variable and may affect the SF-OSDA performance and computational cost. So, knowing the minimum amount of target data the model requires to be competitive with the baselines regarding runtime and performance is crucial. In the following, we evaluate Proto SF-OSDA under these two aspects. All experiments are conducted for Kather-16 → Kather-19 adaptation.

TABLE 3: **Ablation on types of distillation loss** for adapting source model f_s to target domain (Kather-16 → Kather-19) using pseudo-logits of Eq. 8. We compare the performance in terms of ACC and AUC metrics of the target training pseudo-labels with that of the adapted model f_t trained with cross-entropy (CE), SHOT, l_1, and l_2 on the target test set. Red denotes severe degradation while cyan denotes minor change.

Split	Metric	target pseudo-labels	Distillation loss			
			CE	SHOT	l_1	l_2
1	ACC	98.1	97.3(-0.8)	96.8(-1.3)	97.4(-0.7)	98.0(-0.1)
	AUC	99.2	80.1(-19.1)	81.2(-18)	98.8(-0.4)	99.3(+0.1)
2	ACC	98.7	98.7(-0.0)	98.5(-0.2)	98.6(-0.1)	98.7(-0.0)
	AUC	92.7	68.2(-25.5)	68.6(-24.1)	87.8(-4.9)	92.2(-0.5)
3	ACC	99.4	98.4(-1.0)	98.1(-1.3)	98.4(-1.0)	99.0(-0.4)
	AUC	99.0	86.2(-12.8)	85.3(-13.7)	98.4(-0.6)	98.8(-0.2)

Source model architecture sensitivity. We assess our method using different source model architectures, including MobileNet-V2, ViT-Ti/16 [44], and ResNet18 [45] following the same training and adaptation procedure for each split. In Fig. 9, we compare Proto SF-OSDA's results against CE+, BN, CSI, and SHOT methods. We empirically show that our approach significantly surpasses all the baselines regarding ACC and AUC metrics for all three model architectures. Moreover, Proto SF-OSDA's performance seems invariant to the model architecture, an essential requirement for practicality.

Comparing performance, runtime, and dataset size. Fig. 10 shows the runtime of Proto SF-OSDA for adapting the source model on various target dataset sizes, along with their ACC and AUC scores. Our method achieves superior performance by using only 3.5k (5%) unlabeled target domain images, outperforming other SF-OSDA baselines that utilize all 70k (100%) target domain images. This shows the practical significance of Proto SF-OSDA in low data regimes, reducing adaptation computational costs. Note that T3PO learns an auxiliary task of augmentation prediction during source training to detect Open Set categories, while BN updates batch-normalization statistics of the source model with target domain images. However, these methods underperform under severe domain shifts.

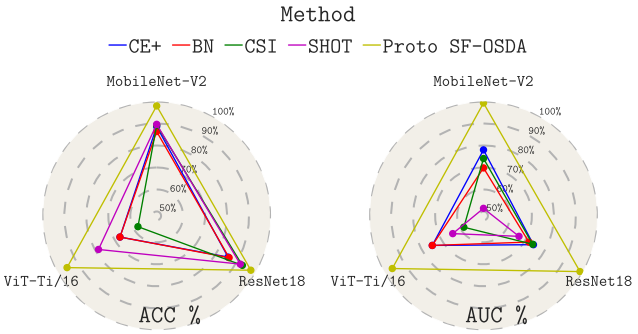

Figure 9: **Model architecture sensitivity** on Kather-16 → Kather-19 adaptation (Split 1). Comparison of three source model architectures: MobileNet-V2, ResNet-18, and ViT-Ti/16 in terms of ACC↑ and AUC↑.

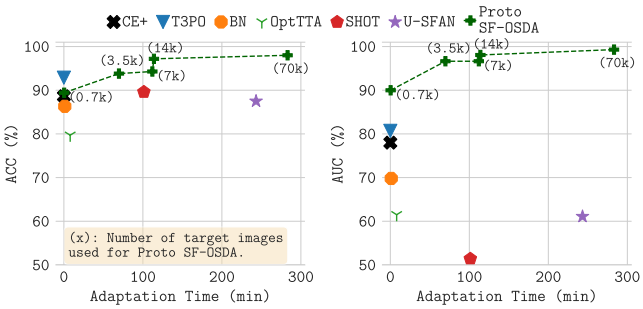

Figure 10: **Comparing performance with runtime and dataset size.** We plot Proto SF-OSDA's performance in terms of ACC/AUC metrics on Kather-16 → Kather-19 (Split 1) adaptation against its runtime and varying amounts of target data along with Open Set detection, and TTA, SF-OSDA methods that utilizes full target dataset (70k images).

6. Conclusion

We presented a novel source-free Open Set domain adaptation method, Proto SF-OSDA, which enhances model robustness to simultaneous semantic and covariate shifts and disregards clinically irrelevant image regions from histological slides. Our key contribution is distilling knowledge from a self-supervised vision transformer trained using our proposed adversarial style augmentations in the target domain. Our method significantly surpasses several SOTA baselines on three colorectal tissue datasets. We conducted several ablation experiments to highlight the importance of individual components of Proto SF-OSDA. In particular, our method performs well even in the low target data regime and with different source model architectures, highlighting its clinical relevance and practicality. As our method is generic, it can be applied to other medical imaging modalities for Open Set domain adaptation in the absence of source data.

References

[1] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, p. e1002730, 2019.

[2] S. Javed, A. Mahmood, M. M. Fraz, N. A. Koohbanani, K. Benes, Y.-W. Tsang, K. Hewitt, D. Epstein, D. Snead, and N. Rajpoot, "Cellular community detection for tissue phenotyping in colorectal cancer histology images," *Medical image analysis*, vol. 63, p. 101696, 2020.

[3] S. Qu, T. Zou, F. Röhrbein, C. Lu, G. Chen, D. Tao, and C. Jiang, "Upcycling models under domain and category shift," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 019–20 028.

[4] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6028–6039.

[5] S. Qu, G. Chen, J. Zhang, Z. Li, W. He, and D. Tao, "Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation," in *European Conference on Computer Vision*. Springer, 2022, pp. 165–182.

[6] Z. Nado, S. Padhy, D. Sculley, A. D'Amour, B. Lakshminarayanan, and J. Snoek, "Evaluating prediction-time batch normalization for robustness under covariate shift," *arXiv preprint arXiv:2006.10963*, 2020.

[7] D. Tomar, G. Vray, J.-P. Thiran, and B. Bozorgtabar, "Opttta: Learnable test-time augmentation for source-free medical image segmentation under domain shift," in *Medical Imaging with Deep Learning*, 2022.

[8] A. Galdran, K. J. Hewitt, N. Ghaffari Laleh, J. N. Kather, G. Carneiro, and M. A. González Ballester, "Test time transform prediction for open set histopathological image recognition," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 263–272.

[9] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=5hLP5JY9S2d

[10] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[11] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *Advances in Neural Information Processing Systems*, 2020.

[12] B. Bozorgtabar, G. Vray, D. Mahapatra, and J.-P. Thiran, "Sood: Self-supervised out-of-distribution detection under domain shift for multi-class colorectal cancer tissue types," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3324–3333.

[13] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[14] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.

[15] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 613–628.

[16] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 813–822.

[17] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[18] J. Linmans, J. van der Laak, and G. Litjens, "Efficient out-of-distribution detection in digital pathology using multi-head convolutional neural networks," in *Medical Imaging with Deep Learning*, 2020.

[19] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *Advances in neural information processing systems*, vol. 31, 2018.

[20] J. N. Kundu, N. Venkat, R. V. Babu *et al.*, "Universal source-free domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4544–4553.

[21] D. Chen, D. Wang, T. Darrell, and S. Ebrahimi, "Contrastive test-time adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 295–305.

[22] D. Tomar, G. Vray, B. Bozorgtabar, and J.-P. Thiran, "Tesla: Test-time self-learning with automatic adversarial augmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 341–20 350.

[23] Q. Wang, O. Fink, L. Van Gool, and D. Dai, "Continual test-time domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7201–7211.

[24] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 922–15 932.

[25] Y. Sun, X. Wang, Z. Liu, J. Miller, A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International conference on machine learning*. PMLR, 2020, pp. 9229–9248.

[26] Y. Liu, P. Kothari, B. Van Delft, B. Bellot-Gurlet, T. Mordan, and A. Alahi, "Ttt++: When does self-supervised test-time training fail or thrive?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 808–21 820, 2021.

[27] Y. Su, X. Xu, and K. Jia, "Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 543–17 555, 2022.

[28] S. Roy, M. Trapp, A. Pilzer, J. Kannala, N. Sebe, E. Ricci, and A. Solin, "Uncertainty-guided source-free domain adaptation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*. Springer, 2022, pp. 537–555.

[29] J. N. Kundu, N. Venkat, A. Revanur, R. V. Babu *et al.*, "Towards inheritable models for open-set domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 376–12 385.

[30] Y. Luo, Z. Wang, Z. Chen, Z. Huang, and M. Baktashmotlagh, "Source-free progressive graph learning for open-set domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[31] Y. Zhang, Z. Wang, and W. He, "Class relationship embedded learning for source-free unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7619–7629.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 296–23 308, 2021.

[34] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[36] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, and F. G. Zöllner, "Multi-class texture analysis in colorectal cancer histology," *Scientific reports*, vol. 6, no. 1, pp. 1–11, 2016.

[37] C. Abbet, L. Studer, A. Fischer, H. Dawson, I. Zlobec, B. Bozorgtabar, and J.-P. Thiran, "Self-rule to multi-adapt: Generalized multi-source feature learning using unsupervised domain adaptation for colorectal cancer tissue detection," *Medical Image Analysis*, vol. 79, p. 102473, 2022.

[38] G. W. van Pelt, T. P. Sandberg, H. Morreau, H. Gelderblom, J. H. J. van Krieken, R. A. Tollenaar, and W. E. Mesker, "The tumour–stroma ratio in colon cancer: the biological role and its prognostic impact," *Histopathology*, vol. 73, no. 2, pp. 197–206, 2018.

[39] C. Abbet, L. Studer, I. Zlobec, and J.-P. Thiran, "Toward automatic tumor-stroma ratio assessment for survival analysis in colorectal cancer," in *Medical Imaging with Deep Learning*, 2022.

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[42] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

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

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[45] 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*, 2016, pp. 770–778.