This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SOoD: Self-Supervised Out-of-Distribution Detection Under Domain Shift for Multi-Class Colorectal Cancer Tissue Types

Behzad Bozorgtabar[1,2,*] Guillaume Vray[1,*] Dwarikanath Mahapatra[3] Jean-Philippe Thiran[1,2]
[1]LTS5, EPFL, Switzerland [2]CIBM, Switzerland [3]Inception Institute of AI, UAE
{behzad.bozorgtabar, firstname.lastname}@epfl.ch {firstname.lastname}@inceptioniai.org

Abstract

The goal of out-of-distribution (OoD) detection is to identify unseen categories of inputs different from those seen during training, which is an important requirement for the safe deployment of deep neural networks in computational pathology. Additionally, to make OoD detection applicable in clinical applications, one may encounter image data distribution shifts. This paper argues that practical OoD detection should handle both semantic shift and data distribution shift simultaneously. We propose a new self-supervised OoD detector for colorectal cancer tissue types based on a clustering scheme. Our work's central tenet benefits from multi-view consistency learning with a supplementary view based on style augmentation to mitigate domain shift. The learned representation is then adapted to minimize images' predictive entropy to segregate in-distribution examples from OoDs on the target data domain. We evaluated our method on two public colorectal tissue types datasets. Our method achieved state-of-the-art OoD detection performance over various self-supervised baselines. The code, data, and models are available at https://github.com/BehzadBozorgtabar/SOoD.

1. Introduction

Colorectal Cancer (CRC) is considered one of the most occurring cancers worldwide, and early-stage CRC diagnosis can significantly improve the chances for therapy of patients [6]. In CRC, the Tumor MicroEnvironment (TME) analysis plays an essential role in cancer grading, and prognostication [23]. Thus, developing automatic tissue phenotyping in Whole Slide Images (WSIs) is of great importance. In recent years, deep learning models have been widely developed for multi-class tissue type classification [24, 38, 2]. While these deep models implicitly assume that the datasets are independent and identically distributed (i.i.d), in practice, collected datasets are typically far from the i.i.d assumption. Histological images present high appearance variability in a real-world scenario due to acquiring data in various conditions, including different scanners or staining procedures. To mitigate this issue, domain adaptation techniques [14, 13, 1, 36] disclose the inference-time data (target domain) to model for adapting the representation from the training data (source domain). Nevertheless, most domain adaptation methods [14, 13] assume a closed-set scenario, where the source and target domains share the same distribution of classes (label set).

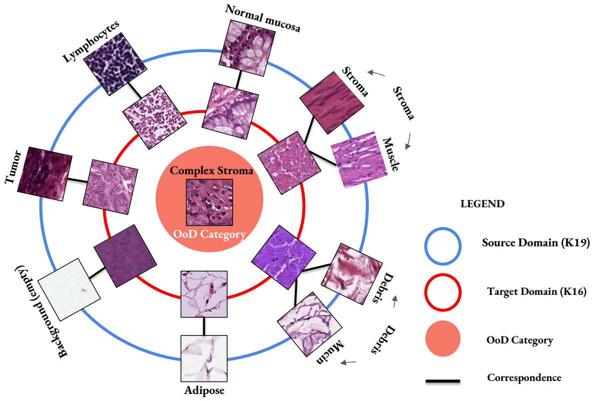

Figure 1: **Motivation of the proposed OoD detection under domain shift.** SOoD seeks to handle both data distribution shift and semantic shift. Histological images of different tissue types present high appearance variability between the source domain (K19) and the target domain (K16), and the target domain contains an additional unknown class (complex stroma).

In clinical routine, a model is often exposed to new data with unknown categories, e.g., tissues from specific cancer subtypes. Thus, making a model robust to the presence of out-of-distribution (OoD) samples and sidestep potentially inaccurate predictions is crucial for the model's safe deployment. Although the task of OoD detection has seen considerable progress [17, 34, 40], developing practical OoD

[*] denotes equal contribution

detection for computational pathology has been a particularly challenging problem for two reasons. *First*, deep neural networks (DNNs) often make overconfident predictions to unknown inputs [31]. *Second*, due to the domain discrepancy mentioned earlier, OoD detectors may mistakenly detect a test sample from known categories but have a different style/domain as an OoD. Fig. 1 shows the correspondence of different tissue types across two CRC datasets, Kather-19 (K19) [24] and Kather-16 (K16) [25], as the source domain and target domain. Histological images present high appearance variability between two data domains, and the target domain contains an additional unknown class (complex stroma).

To address the limitations of current OoD detection methods, we propose **SOoD**, short for **S**elf-supervised **O**ut-**o**f-**D**istribution detection under domain shift for multi-class colorectal cancer tissue types, a new self-supervised OoD detector to mitigate both semantic shift and data distribution shift. We illustrate the pre-training stage of SOoD in Fig. 2 and present a pseudo-code implementation in Algorithm 1. In summary, we highlight the contributions below:

- Our method (SOoD) is the first work to consider the problem of multi-class OoD detection under domain shift for clinical applications to the best of our knowledge;
- The proposed self-supervised OoD method builds upon multi-view consistency paradigm with complementary style augmentation to mitigate domain shift, as opposed to current OoD detections, which focus on a single image domain;
- We propose a new self-training scheme for OoD detection via minimizing images' predictive entropy of unlabeled images to segregate in-distribution examples from OoDs on the target data domain. Our method does not require OoD samples during training and is capable of working with unlabeled source datasets alleviating costly annotations;
- Experimental results show consistent improvement of proposed OoD detection performance over state-of-the-art (SOTA) self-supervised methods [8, 7, 42, 3] on two hematoxylin & eosin (H&E) stained CRC tissue types datasets [24, 25].

2. Related Work

The problem of OoD detection has seen considerable progress in computer vision and medical image analysis, as OoD detection is crucial for the safe deployment of deep learning systems. The related work in this area is sizable. Thus, we mainly focus on the recent deep learning-based methods in supervised [15] and unsupervised [44] settings.

Most of the current OoD detection solutions presume access to the OoD datasets during training [22, 47] or validation steps [27, 26, 34] that are not well suited for the general use of OoD detection in real-world applications. Some interesting methods [16, 27, 26] benefit from adversarial samples via perturbation of the training samples to improve the robustness of their network, which results in higher training time and suboptimal solutions.

On the other hand, recent methods [17, 34, 40] rely on generative or reconstruction-based training schemes [39, 49], deep one-class classifiers [35, 29], and, more recently, self-supervised approaches [15, 5, 30, 3]. Overall, the underlying rationale behind those methods is modeling the representation of in-distribution data either using a one-class [35] or multi-class setup [4], and then a detection function is usually defined to detect OoDs. However, most previous OoD detection methods assume that the training and test data would follow a similar distribution (style/domain). This assumption can negatively impact OoD detection as OoD detectors may erroneously detect a test sample from known classes but have a different style as an OoD class. A possible solution would be to use additional data from the new target domain and formulate this problem as open-set domain adaptation [37, 32], where the source domain contains in-distribution labeled data and the target domain contains novel classes in addition to the classes present in the source domain. Nevertheless, it would require labeled source data and costly annotations by domain experts. Recent studies have shown that contrastive training [8, 19, 9] significantly improves OoD detection [45, 42]. These methods attempt to learn representation based on attracting similar views of a sample and repelling disagreeing views from each other. However, current contrastive training methods are incentivized to learn features from a single image domain.

3. Method

We start by motivating our approach before explaining the methodological details. The main goal of OoD detection under domain shift is to learn domain-invariant representation between one specific domain (i.e., source domain) and a testing domain (i.e., target domain) so that an OoD model can robustly leverage such invariances to a new unlabeled target domain. The domain invariance is often ignored or not formulated in previous OoD detection methods. As a result, *OoD detectors may mistakenly detect a new test example from known classes but have a different style as an OoD class.*

This problem setting is also different from typical unsupervised domain adaptation (UDA) approaches as a new target domain contains an additional unknown class. Besides, unlike UDA methods, we formulate our proposed OoD to deal with the unlabeled source data, which is highly demanded in practical applications. In this work, we revisit current state-of-the-art contrastive learning-based self-supervised methods [10, 7, 8, 18] for the OoD task using only positive pair samples. In particular, we extend the two-view consistency learning paradigm based on self-augmentation to a multi-view version with style augmentation as the new complementary view.

The objective of the pre-training stage is to simultaneously learn *domain-invariant features* and *consistent cluster assignments* between multiple views of the same tissue image in an entirely unsupervised setting (Sect. 3.1). The learned representation is then adapted using a *self-training* scheme on the unlabeled target domain images. Typically, one can use the most probable cluster predicted by the network as pseudo labels. Since the pseudo labels are often noisy, we propose to segregate *known* and *unknown* samples using the entropy of the clustering output and opt for only highly confident (lower entropy) target images for pseudo labeling. More specifically, we perform entropy minimization on selected unlabeled target images w.r.t source prototypes from (Sect. 3.1) to segregate known categories from OoDs (Sect. 3.2). Finally, we define the OoD function to detect OoDs.

3.1. Multi-View Consistency

We revisit the recent self-supervised clustering scheme [7], which clusters the data while imposing an agreement between cluster assignments obtained from different augmentations of the same image. Specifically, we address the limitation of the current self-augmentation consistency learning paradigm in the presence of data domain shift. To do so, we extend typical two-view consistency learning to a multi-view version with style augmentation of the target domain as the new complementary view. Each source domain image x_s is transformed into two in-domain augmented views, including *weakly augmented view* x_{sw} and *heavily augmented view* x_{sh}, and the encoder output of the weakly augmented view provides a pseudo label for the predictions on heavily augmented view. Furthermore, we add an additional view based on *style augmentation* x_{style} to make the model robust against domain shift, especially in the absence of labeled data. As for the style augmentations, weakly augmented images are mapped from the source domain to the target domain via a pre-trained CycleGAN model [48]. Such a new view makes the model invariant to the image style by further covering the target data distributions and adding the regularization effect through multi-view consistency learning.

Algorithm 1: SOoD PyTorch pseudocode w/o multi-crop (pre-training stage).

```
Input: $\mathcal{S}$: unlabeled or partially labeled source samples, $trslt$:
       pre-trained style transformer on $\mathcal{S}$ and unlabeled target
       samples $\mathcal{T}$, $f_\theta$: encoder network
Output: updated $f_\theta$
Parameter : tp: temperature, $\lambda_1$, $\lambda_2$: weights for the loss
       terms, sinkhorn: Sinkhorn-Knopp function
for x in loader do // load a minibatch with n samples from $\mathcal{S}$
    x_sw = weak_augment(x)// augmented views
    x_sh = heavy_augment(x)
    x_style = $trslt$(x_sw)
    scores_sw, scores_sh, scores_style = $f_\theta$(x_sw), $f_\theta$(x_sh),
      $f_\theta$(x_style)// output n-by-K
    pseudo_sw, pseudo_sh, pseudo_style = sinkhorn(scores_sw),
      sinkhorn(scores_sh), sinkhorn(scores_style)// apply
      sinkhorn to generate pseudo label
    $\ell_{heavy}$ = H(pseudo_sw, scores_sh)/2 + H(pseudo_sh,
      scores_sw)/2
    $\ell_{style}$ = H(pseudo_sw, scores_style)/2 + H(pseudo_style,
      scores_sw)/2
    $\ell_{mv} = \lambda_1 \ell_{heavy} + \lambda_2 \ell_{style}$
    $\ell_{mv}$.backward() // back-propagate
    update($f_\theta$) // encoder update
def H(pseudo, score):
    pseudo = pseudo.detach()// stop gradient
    pred = softmax(score / tp, dim=1)
    return-(pseudo * log(pred)).sum(dim=1).mean()
```

More precisely, we apply a non-linear mapping f_θ to the multi-view augmented images to match their representations to K dimensional features. The non-linear encoder f_θ includes the convolutional neural network (CNN) backbone followed by a 2-layer MLP network. Given an image x from one of the three different augmentations of the input source image, we compute its cluster assignments (codes) by matching its feature representations to a set of K trainable prototypes $\{c_1, \cdots, c_K\}$. These soft assignments are in the form of probability distributions over K dimensions. Then, the probability P is obtained by normalizing the output of the encoder f_θ with a softmax function:

$$P(x)^{(i)} = \frac{\exp\left(f_\theta(x)^{(i)} / \tau\right)}{\sum_{k=1}^{K} \exp\left(f_\theta(x)^{(k)} / \tau\right)} \tag{1}$$

where τ is a temperature parameter [46]. As in [7], features before last linear layer of f_θ and prototypes are ℓ_2 normalized. We optimize the multi-view consistency loss ℓ_{mv} w.r.t. the parameters of the encoder θ. Thus, the encoder output of the weakly augmented view provides a pseudo label for predictions of two other augmented views from the source image based on heavy and style augmentation through the cross-entropy losses ℓ_{heavy} and ℓ_{style},

$$\min_\theta \ell_{mv} = \lambda_1 \ell_{heavy} + \lambda_2 \ell_{style} \tag{2}$$

where λ's denote the weights for the heavily augmented view loss ℓ_{heavy} and style augmented view loss ℓ_{style}. The pseudo label is obtained by applying the iterative *Sinkhorn-*

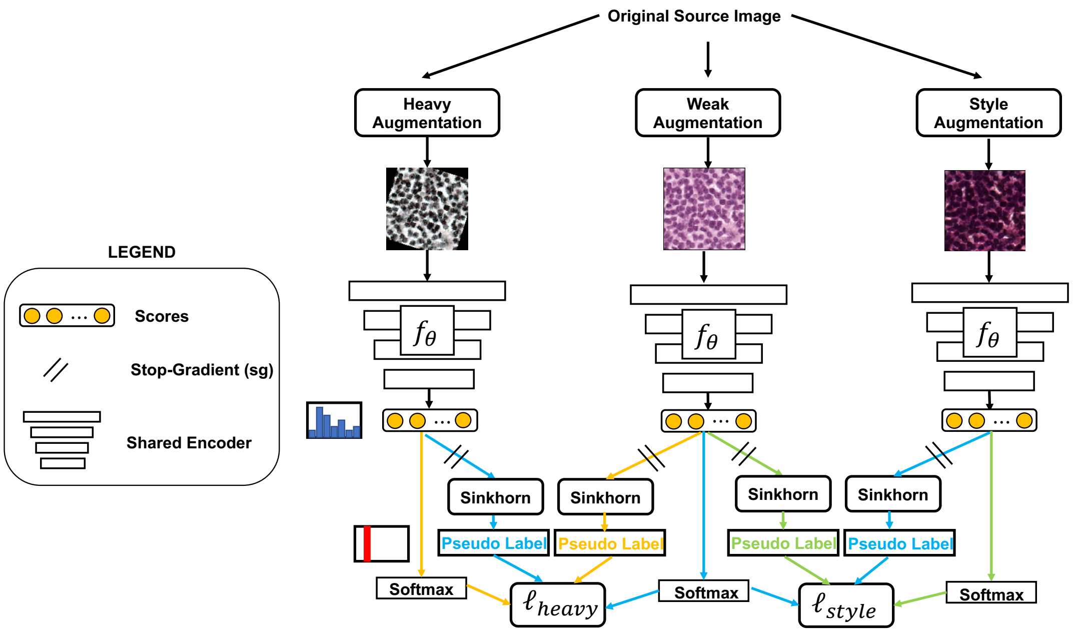

Figure 2: **The overview of the pre-training stage of SOoD**. Augmented views for input source images are generated, and the multi-view consistency loss is optimized. The encoder output of the weakly augmented view provides a pseudo label for predictions of two other augmented views from the source image based on heavy and style augmentation.

Knopp algorithm [12] on the output of the encoder f_θ to select all prototypes the same amount of time.

Details of the loss terms. We first describe the loss term ℓ_{heavy} for the heavily augmented view, and a similar formula holds for the loss term of the style augmented view ℓ_{style}. For the heavily augmented view, we use RandAugment [11] that mainly deals with color intensity and geometrical transformations. We first compute for an unlabeled weakly augmented view its pseudo label $\hat{P}\left(x_{sw}\right) \in \{1, \cdots, K\}$ w.r.t the K prototypes. This is achieved by first applying a stop-gradient (sg) operator on the encoder output and then using iterative *Sinkhorn-Knopp* algorithm [12], $\hat{P}\left(x_{sw}\right) = \text{sinkhorn}(\text{sg}\left(P\left(x_{sw}\right)\right))$. Then, we optimize the encoder to match the heavily augmented view prediction $P\left(x_{sh}\right)$ to the pseudo label $\hat{P}\left(x_{sw}\right)$ using the cross-entropy loss. In practice, we additionally benefit from a multi-crop data augmentation strategy [7] such that from a given image, we generate a set V of different positive views. This set contains two anchor views and several local image crops of smaller resolution. The predictions of all crops are attracted to the anchor views to further improve the quality of the learned embeddings. We minimize the loss ℓ_{heavy} with stochastic gradient descent:

$$\ell_{heavy} = \min_{\theta} \sum_{x_{sw}\in\left\{x_1^a, x_2^a\right\}} \sum_{x'_{sh}\in V, x'_{sh}\neq x_{sw}} H\left(\hat{P}\left(x_{sw}\right), P\left(x'_{sh}\right)\right) \tag{3}$$

where $H\left(a, b\right) = -a \log b$, and x_1^a and x_2^a denote the anchor views. Similar formula holds for ℓ_{style} to align the style augmented view prediction $P\left(x_{style}\right)$ to the pseudo label $\hat{P}\left(x_{sw}\right)$ using the cross-entropy loss:

$$\ell_{style} = \min_{\theta} \sum_{x_{sw}\in\left\{x_1^a, x_2^a\right\}} \sum_{x'_{style}\in V, x'_{style}\neq x_{sw}} H\left(\hat{P}\left(x_{sw}\right), P\left(x'_{style}\right)\right) \tag{4}$$

The style augmented view loss ℓ_{style} complements heavily augmented view loss ℓ_{heavy} by making the encoder robust to style variation present in the target domain. Following [18], we use a symmetrized loss for both loss terms (Eq. 3 & Eq. 4) as symmetrization helps boost accuracy (see Algorithm 1).

3.2. Self-Training via Entropy Minimization

We incorporate an additional self-training criterion on the target domain into our model to further facilitate OoD detection. For this purpose, the pre-trained encoder f_θ from the previous step is applied on unlabeled target images to generate the pseudo-label for target samples, which are then used to fine-tune the encoder. Since we are not using label information on the target domain, we use the entropy of the cluster assignment to draw a boundary between in-distribution and OoDs such that we expect that the entropy of OoDs is larger than entropy for the in-distribution samples. To determine the optimal threshold for the entropy,

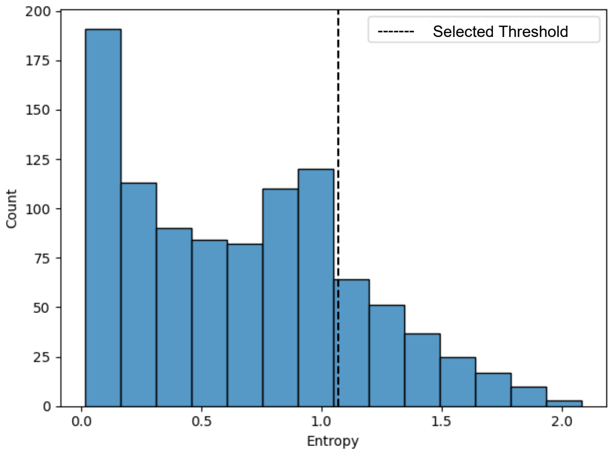

Figure 3: **The entropy histogram of cluster assignments** for the style augmented images. We set the threshold ρ to 1.07 such that the entropy of $\simeq$ 80% of training examples will be lower than ρ. This threshold is used to select highly confident target examples for pseudo labeling.

we first compute the entropy of the style augmented images used for training and select a threshold ρ such that the majority of the style augmented images ($\simeq$80%)[1] have an entropy lower than ρ. Then we apply this threshold to the unlabeled images from the target domain to select highly confident samples for pseudo labeling. We perform an analysis of ρ in Fig. 3. For the self-training, we perform predictive entropy minimization on pseudo-labeled target data to make them tighter clustered around the source prototypes $\{c_1, \cdots, c_K\}$. This increases the confidence of cluster predictions and identifies OoDs if they have different characteristics compared to in-distribution samples. The prototypes are kept fixed during self-training, and only the parameters of f_θ are updated. We minimize the entropy loss for the self-training step ℓ_{st} as follows:

$$\ell_{st} = \mathbb{E}_{x_t \sim \mathcal{T}} \left[\sum_{k=1}^{K} -P\left(x_t\right)^{(k)} \log P\left(x_t\right)^{(k)} \right] \tag{5}$$

where $P\left(x_t\right)^{(k)}$ is the probability obtained by the encoder shows unlabeled target sample x_t matches with cluster prototype c_k.

Inference: At inference time, a test image x_{test} is passed through the trained encoder f_θ to obtain its feature representation $v_{test} = f_\theta\left(x_{test}\right)$. v_{test} is then compared with the top M similar features $\{v_m\}$ of the target domain's training samples based on the cosine similarity. An OoD detection score $\mathcal{S}\left(\cdot,\cdot\right)$ is computed as follows:

$$\mathcal{S}\left(v_{test}; \{v_m\}\right) := -\frac{1}{M} \sum_{m=1}^{M} sim(v_m, v_{test}) \tag{6}$$

where $sim(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^T \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$ and $\mathcal{S}\left(\cdot,\cdot\right)$ is normalized using the maximum and minimum scores of the set such that $\mathcal{S}\left(\cdot,\cdot\right) \in [0, 1]$. Intuitively, the scores of OoD samples should be larger than the scores from in-distribution ones.

4. Experiments

Datasets and Evaluation Metrics. We evaluate SOoD on two H&E stained publicly available CRC datasets, Kather-19 (K19) [24] and Kather-16 (K16) [25], as the source domain and target domain. K16 dataset contains 5,000 images patches of 150 × 150 pixels each ($74\mu m \times 74\mu m$) from H&E WSIs, while K19 dataset contains 100,000 H&E stained patches at (0.5 μm/pixel). There is a data distribution shift across two image domains together with a semantic shift of tissue phenotypes. Incorporating expert pathologists' feedback [1], we group debris/mucus and stroma/muscle as debris and stroma, respectively, to correspond between the two datasets. As a result, we end up with seven tissue categories shared between two domains, including (tumor, stroma, lymphocytes, debris, normal mucosa, adipose, and background or empty class). The target domain contains an additional tissue category of complex stroma that is not present in the source domain, and we consider this tissue type as OoD class. In total, we end up with 11,495 training images (7,995 from the source domain and 3,500 from the target domain) without using OoDs. For the validation set, we use 997 images from the source domain (pre-training and self-training), 621 images from the target domain for t-SNE visualization purpose. For the test set, we use 879 images from the target domain, including 438 OoD images. The rest of the test images are equally distributed between seven in-distribution classes. We use OoD detection metrics: area under the ROC curve (AUC) and area under the precision-recall curve (AUPRC) and present mean ± std on the test set for all experiments over three runs. Our experiments follow the setting for multi-crop using two anchor views at resolution 144 × 144 pixels and multiple small crops (local views) of resolution 96 × 96 pixels.

Implementation Details. Our implementations are based on PyTorch 1.9 [33]. We adopt the ResNet18 [20] as the backbone network for SOoD. All networks are trained using SGD optimizer ($momentum = 0.9$), with a weight decay of $1e-6$ and a learning rate of 0.06. A cosine scheduler is used during the training. A hyper-parameter search was conducted to find the optimal batch size (64), τ (0.1), prototypes (16), λ_1 (1) and λ_2 (1). Also, we found

[1] This ratio is determined based on the distribution of unknown samples.

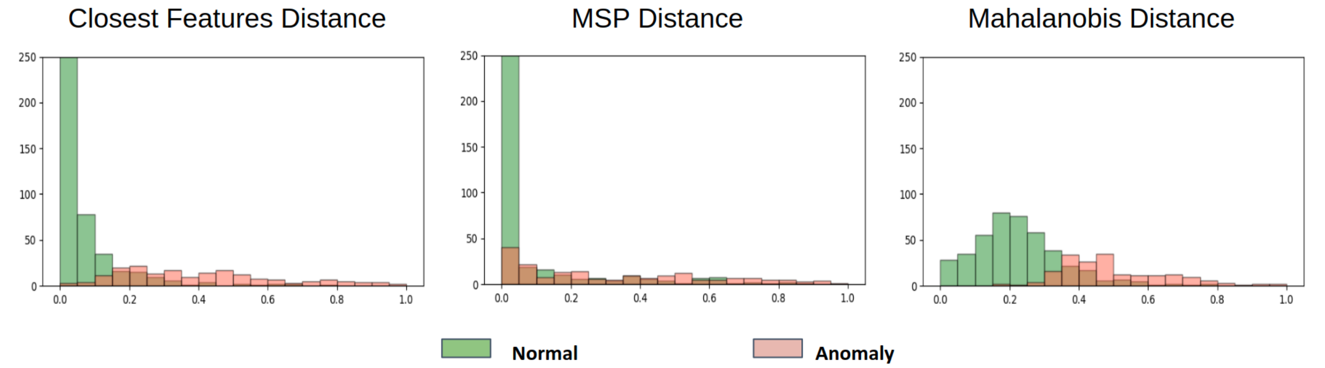

Figure 4: **The histograms of OoD detection scores** for in-distribution (normal) and OoDs (anomalies) on the target set (K16). We compare our OoD detection score (left) to other anomaly scores (middle-right). Our OoD detection score clearly discriminates in-distribution and OoD test images.

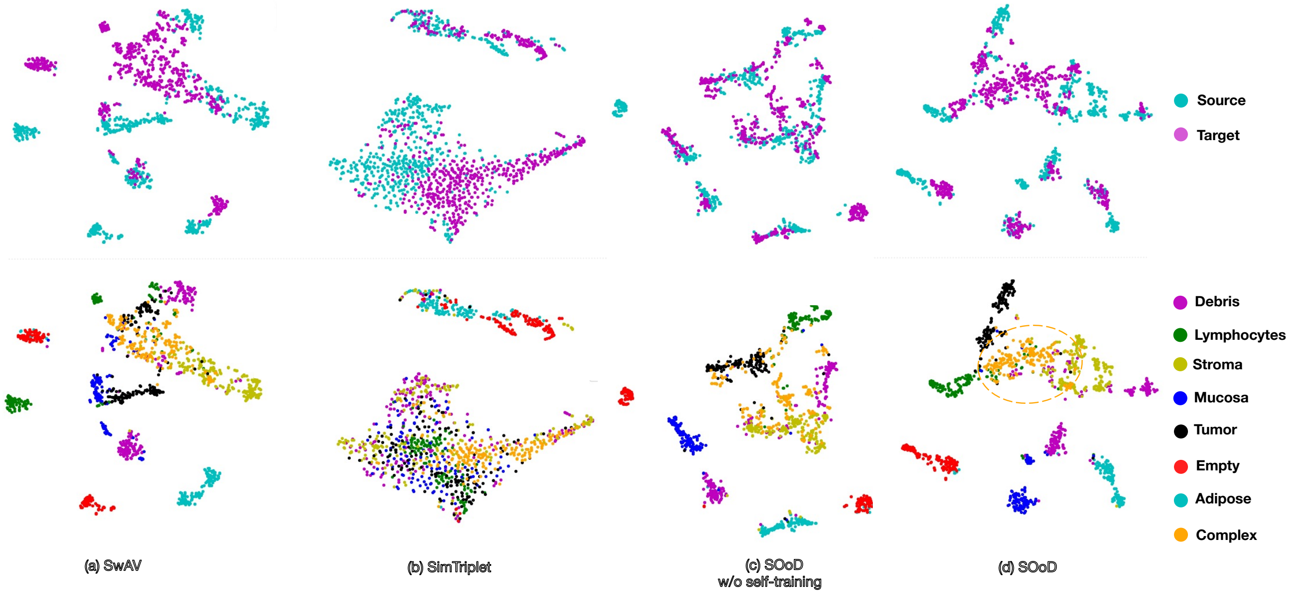

Figure 5: **The t-SNE [43] visualization** of the feature representations extracted by the encoder trained on the source (K19) and target sets (K16) for domain alignment (top row) and the different classes' representations (bottom row). We compare our method (c-d) to other SOTA self-supervised methods (a-b). Further fine-tuned with self-training, our model learns domain-invariant representation and can separate OoDs from in-distribution test samples (highlighted with dashed orange contour).

the optimal number of nearest neighbors for k-NN and $M \in \{5, 10, 20, 50\}$ in Eq. 6 and set it to 10 for all baselines. We conduct ablation studies for the chosen λ's and the number of prototypes. We apply random resize crops and horizontal flips augmentation for the weakly augmented view. Besides, we use RandAugment [11] for obtaining the heavily augmented view. Our model has been pre-trained for 300 epochs using ℓ_{mv} and then fine-tuned by minimizing ℓ_{st} for additional 20 epochs with a lower learning rate of 0.001. For a fair comparison with [7], we set the Sinkhorn regularization parameter ϵ to 0.05 and use three iterations for all runs. To ensure consistent and comparable comparisons, we use the same experimental setup for all baselines.

Comparison with SOTA Methods. Since no prior work has been done for our specific setup, we provide our baselines for OoD detection under domain shift based on state-of-the-art self-supervised learning methods. We first validate the SOoD framework used in this study with state-

* $p < 0.001$; a bilateral Welch t-test with respect to the top result.
[†] We compute the OoD detection score before self-training, using only the pre-trained model.
[‡] We compute the OoD detection score after self-training.

(A)	Source Only		Target Only		Both Domains	
Method	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
SimCLR [9]	72.79 ± 1.64	66.84 ± 1.93	87.05 ± 1.24	81.15 ± 2.10	88.75 ± 1.14	83.51 ± 1.97
SwAV [7]†	70.70 ± 1.76	60.80 ± 1.80	84.42 ± 1.32	76.73 ± 2.13	76.84 ± 1.66	69.68 ± 2.17
SwAV [7]‡	78.50 ± 1.55	72.57 ± 2.09	87.34 ± 1.25	80.81 ± 2.12	85.73 ± 1.28	81.96 ± 1.88
CDMSAD [41]	72.85 ± 1.65	66.85 ± 1.96	84.03 ± 1.34	78.64 ± 2.03	69.34 ± 1.86	58.47 ± 1.70

(B)	CSI [42]	GOAD [3]	SimTriplet [28]	SwAV [7]	DINO [8]	Our model w/o self-training†	**SOoD‡**
AUROC	89.64 ± 1.95	70.56 ± 2.42	69.53 ± 1.32	67.58 ± 2.77	52.56 ± 0.43	88.38 ± 1.30	**92.77 ± 0.48**
AUPRC	86.32 ± 2.21	63.11 ± 1.84	60.01 ± 1.42	60.44 ± 2.81	46.52 ± 0.29	80.43 ± 2.84	**90.90 ± 1.00**

(C)	Sup. Tr.-100%	Sup. Src.-100%	SwAV-100%	DINO-100%	**SOoD-100%**	**SOoD-20%**	SOoD-10%	SOoD-1%
Linear (ACC)	78.31 ± 5.98	65.13 ± 3.57	48.83 ± 1.83	42.03 ± 8.51	**73.24 ± 0.39**	**73.39 ± 0.69**	73.24 ± 0.82	62.59 ± 1.42
Linear (F1)	78.27 ± 5.86	62.89 ± 1.15	45.67 ± 2.31	37.17 ± 8.49	**69.88 ± 0.92**	**70.29 ± 0.72**	70.22 ± 0.92	60.29 ± 0.94
k-NN (ACC)	77.48 ± 1.61	41.50 ± 3.84	41.72	32.20	**83.45**	-	-	-
k-NN (F1)	78.01 ± 1.48	37.28 ± 3.97	37.11	28.46	**83.25**	-	-	-

Table 1: Evaluation of the proposed method and baselines. Section-wise best scores are in **Bold**.

of-the-art OoD detection methods, including contrastive training-based methods and self-supervised approaches. We also evaluate the representation quality by following common practice in self-supervised learning with a linear classifier on top of frozen features from the pre-training stage. The baselines include:

- state-of-the-art contrastive training based OoD detection methods, including CSI [42], open-set OoD detection (GOAD) [3], and CDMSAD [41],
- self-supervised learning-based methods [9, 7] pre-trained on *source domain*, *target domain*, and *both domains* [28, 8, 7], respectively,
- supervised linear classification on frozen features from different self-supervised methods [8, 7].

Results and Discussion. In Table 1 A, we first report the results of state-of-the-art self-supervised methods [9, 7, 41] using the same backbone architecture as in SOoD, a ResNet18 on K16 and K19 datasets. Similar to SOoD, these methods are concerned with the scenarios where the source data are not labeled [9, 7] except for [41], where the source data is partially labeled. We use the same OoD detection score based on the closest feature distance for all the baselines for a fair comparison. The contrastive training-based methods are strong baselines [9, 7, 41] and can generalize reasonably well on the target data with the model pre-trained on the *source data* only, *target data* only, or *both domains*. The models that take advantage of the target distribution gain superior performance improvement than the same models trained on the source data. In addition, combining both data domains for training can further improve performance.

In Table 1 B, regarding the models trained on *both domains*, the GOAD method [3] often incorrectly detects a known class of target domain as an OoD due to the domain shift. CSI [42] benefits from contrastive learning to contrast each image with distributionally-shifted augmentations of itself. The methods in [7, 8] are based on a clustering scheme, and we use the same self-training approach as in SOoD for a fair comparison. SimTriplet [28] is the only method that incorporates multiple instances, but it is not formulated to address domain shift. For the input views of [28], we use the same three augmented views as in SOoD. Unlike these baselines, our method learns generalizable semantic properties in the feature space via clustering. Our designed domain-invariant formulation gains huge improvements under domain shift, and our results outperform other state-of-the-art self-supervised methods. Additionally, self-training used in our method further enhances the OoD detection performance. In Table 1 C, we also evaluate the quality of frozen features from the pre-training stage (pre-trained with ℓ_{mv}) via training a linear classifier on the frozen features. The objective is to show the effectiveness of SOoD to classify in-distribution target images correctly. Furthermore, we use a nearest neighbors classifier (k-NN) without any finetuning to vote for the label of in-distribution test images from the target domain. In Table 1, we report average F1 and accuracy (ACC) scores for seven in-distribution classes (all) using both schemes. We use the fully-supervised trained model (ResNet18) by utilizing all labeled style augmented images as the upper bound (Sup. Tr.-100%)[2]. Self-supervised features from pre-

[2] One can consider a trained, supervised model on fully labeled real target images as the upper bound, but this setting is not realistic as we do not use label information from the target domain.

	K Sensitivity		Loss Weights Sensitivity				Final Model (**SOoD**)
Metric	K=8	K=24	$\ell_1 = 3 \vert \ell_2 = 1$	$\ell_1 = 1 \vert \ell_2 = 3$	w/o ℓ_{heavy}	w/o ℓ_{style}	K=16, $\ell_1 = 1 \vert \ell_2 = 1$
AUROC (A)	82.52 ± 1.69	86.70 ± 4.33	85.18 ± 2.12	82.76 ± 0.86	84.63 ± 0.43	83.95 ± 1.85	**88.38 ± 1.30**
AUROC (B)	88.85 ± 0.57	89.24 ± 0.40	90.77 ± 0.47	90.99 ± 0.28	91.60 ± 0.59	85.87 ± 3.25	**92.77 ± 0.48**
AUPRC (A)	71.44 ± 2.03	79.60 ± 7.56	76.41 ± 3.22	71.80 ± 0.25	75.76 ± 1.07	75.63 ± 2.40	**80.43 ± 2.84**
AUPRC (B)	85.30 ± 0.22	84.71 ± 0.43	87.95 ± 0.33	88.64 ± 0.39	90.62 ± 0.54	84.56 ± 2.26	**90.90 ± 1.00**

Table 2: **Ablation studies** for the different number of prototypes K and loss weight values. We evaluate the models for both before self-training (A) and after the self-training stage (B).

Metric	Color Jittering	**SOoD**
AUROC (A)	82.44 ± 1.20	88.38 ± 1.30
AUROC (B)	88.58 ± 0.73	92.77 ± 0.48
AUPRC (A)	75.23 ± 1.12	80.43 ± 2.84
AUPRC (B)	86.54 ± 0.57	90.90 ± 1.00

Table 3: **Ablation studies** for different augmentation techniques. We evaluate the models for both before self-training (A) and after the self-training stage (B).

Checkpoint	AUROC	AUPRC
	Mahalanobis Distance [26]	
Before Self-Training	79.08 ± 0.98	69.13 ± 1.95
After Self-Training	92.36 ± 0.44	90.22 ± 0.74
	MSP Distance [21]	
Before Self-Training	68.79 ± 3.39	64.56 ± 4.13
After Self-Training	83.66 ± 0.64	79.05 ± 0.85
	Closest Features Distance	
Before Self-Training	**88.38 ± 1.30**	**80.43 ± 2.84**
After Self-Training	**92.77 ± 0.48**	**90.90 ± 1.00**

Table 4: The evaluation of the proposed method with different **OoD detection techniques** before and after self-training.

trained SOoD perform particularly well with either learning a linear classifier or k-NN and surpass SOTA self-supervised methods and supervised baseline. For example, SOoD trained with 10% labeled source data outperforms the fully supervised model trained from scratch on the source domain (Sup. Src.-100%), reducing the gap with full-label training (Sup. Tr.-100%). Finally, compared to [7, 28], the t-SNE [43] visualization of extracted features from the encoder f_θ shows a better alignment of the source and target domains, representations of the known classes, and a better separation of OoDs (see Fig. 5).

Ablations. We provide ablation studies to analyze the key factors that lead to the success of SOoD. These ablations concerning various aspects of SOoD's design, including loss terms, loss weight values, number of prototypes, augmentation techniques, and OoD detection scores (see Table 2). We sweep over a different number of prototypes (16-24) and find that our method is not very sensitive to the number of prototypes, but using fewer prototypes ($< 2 \times$ classes) leads to performance degradation. As argued, the model trained with additional style augmented view achieved a significant performance boost compared to baselines without this complementary view (w/o ℓ_{style}). The sensitivity test for the loss weight values also shows each loss term for augmented views is equally important, improving the regularization effect of multi-view learning. Table 3 compares the OoD detection performance of our style augmentation with other augmentation (color jittering) and shows that the performance is significantly improved as we learn style-invariant representation. For all ablation experiments in Table 2 and Table 3, self-training the model on the target domain yields a better separation of OoDs from known classes and higher accuracy than the model trained only with optimizing ℓ_{mv}. This is also indicated by the t-SNE [43] visualization in Fig. 5. Finally, we compare our proposed OoD detection score with popular techniques (see Table 4), including Maximum over softmax probabilities (MSP) [21] and Mahalanobis distance [26]. The comparison demonstrates that OoD detection in these baselines might be failing due to the semantic ambiguity of some tissue categories, while ours achieves superior performance.

5. Conclusion

Our method is the first self-supervised OoD detection for CRC tissue types under domain shift in a zero-labeled data regime, yielding a more realistic and practical setting and alleviating costly annotations. It is also critical to safely deploy DNNs in computational pathology to generalize to a new clinical site with new categories not presented in a source dataset. We show that our designed multi-view consistency learning together with a self-training scheme gains substantial performance improvements in both OoD detection and classification of in-distribution samples compared to SOTA self-supervised methods. SOoD can be easily adjusted to be applied to different organs and histology tasks. In future work, we plan to design new formulations to improve the accuracy of pseudo labels and the generalizability of our method to unseen datasets.

References

[1] Christian Abbet, Linda Studer, Andreas Fischer, Heather Dawson, Inti Zlobec, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-rule to adapt: Learning generalized features from sparsely-labeled data using unsupervised domain adaptation for colorectal cancer tissue phenotyping. In *Medical Imaging with Deep Learning*, 2021. 1, 5

[2] Babak Ehteshami Bejnordi, Maeve Mullooly, Ruth M Pfeiffer, Shaoqi Fan, Pamela M Vacek, Donald L Weaver, Sally Herschorn, Louise A Brinton, Bram van Ginneken, Nico Karssemeijer, et al. Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Modern Pathology*, 31(10):1502–1512, 2018. 1

[3] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2019. 2, 7

[4] Supritam Bhattacharjee, Devraj Mandal, and Soma Biswas. Multi-class novelty detection using mix-up technique. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1400–1409, 2020. 2

[5] Behzad Bozorgtabar, Dwarikanath Mahapatra, Guillaume Vray, and Jean-Philippe Thiran. SALAD: Self-supervised aggregation learning for anomaly detection on x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 468–478. Springer, 2020. 2

[6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018. 1

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. 2020. 2, 3, 4, 6, 7, 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2, 3, 7

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2, 7

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 4, 6

[12] Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NIPS*, volume 2, page 4, 2013. 4

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 1

[15] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9781–9791, 2018. 2

[16] Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. DROCC: Deep robust one-class classification. In *International Conference on Machine Learning*, pages 3711–3721. PMLR, 2020. 2

[17] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019. 1, 2

[18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3, 4

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 8

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019. 2

[23] Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Medical image analysis*, 63:101696, 2020. 1

[24] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019. 1, 2, 5

[25] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexan-

der Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6(1):1–11, 2016. 2, 5

[26] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 2, 8

[27] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 2

[28] Quan Liu, Peter C Louis, Yuzhe Lu, Aadarsh Jha, Mengyang Zhao, Ruining Deng, Tianyuan Yao, Joseph T Roland, Haichun Yang, Shilin Zhao, et al. Simtriplet: Simple triplet representation learning with a single gpu. *arXiv preprint arXiv:2103.05585*, 2021. 7, 8

[29] Philipp Liznerski, Lukas Ruff, Robert A Vandermeulen, Billy Joe Franks, Marius Kloft, and Klaus-Robert Müller. Explainable deep one-class classification. *arXiv preprint arXiv:2007.01760*, 2020. 2

[30] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5216–5223, 2020. 2

[31] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 2

[32] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 754–763, 2017. 2

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019. 5

[34] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32:14707–14718, 2019. 1, 2

[35] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402, 2018. 2

[36] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 1

[37] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–168, 2018. 2

[38] Can Taylan Sari and Cigdem Gunduz-Demir. Unsupervised feature extraction via deep learning for histopathological classification of colon tissue images. *IEEE transactions on medical imaging*, 38(5):1139–1149, 2018. 1

[39] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 2

[40] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2019. 1, 2

[41] Antoine Spahr, Behzad Bozorgtabar, and Jean-Philippe Thiran. Self-taught semi-supervised anomaly detection on upper limb x-rays. *arXiv preprint arXiv:2102.09895*, 2021. 7

[42] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020. 2, 7

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6, 8

[44] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[45] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020. 2

[46] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[47] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9518–9526, 2019. 2

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[49] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018. 2