Positive-unlabeled learning for open set domain adaptation

Positive-Unlabeled Learning for Open Set Domain Adaptation

Mohammad Reza Loghmani[a,*], Markus Vincze[a], Tatiana Tommasi[b]

[a] *Vision for Robotics laboratory, ACIN, Technische Universität Wien, Vienna 1040, Austria*
[b] *Italian Institute of Technology and DAUIN, Politecnico di Torino, 10138 Torino, Italy*

Abstract

Open Set Domain Adaptation (OSDA) focuses on bridging the domain gap between a labeled source domain and an unlabeled target domain, while also rejecting target classes that are not present in the source as unknown. The challenges of this task are closely related to those of Positive-Unlabeled (PU) learning where it is essential to discriminate between positive (known) and negative (unknown) class samples in the unlabeled target data. With this newly discovered connection, we leverage the theoretical framework of PU learning for OSDA and, at the same time, we extend PU learning to tackle uneven data distributions. Our method combines domain adversarial learning with a new non-negative risk estimator for PU learning based on self-supervised sample reconstruction. With experiments on digit recognition and object classification, we validate our risk estimator and demonstrate that our approach allows reducing the domain gap without suffering from negative transfer.

Keywords: Computer Vision, Deep Learning, Image Classification, Domain Adaptation, Open Set Recognition, Positive-Unlabelled Learning

1. Introduction

The process of acquiring and annotating a large amount of application-specific data is one of the main ingredients of the current success of deep learning, but it is very costly. An attractive solution is to take advantage of large publicly available datasets, which however may not capture the exact characteristics of the application of interest, leading to poor performance at deployment time. The challenges derived from this difference between training (source) and test (target) samples are subject of active research in machine learning with several applications in the area of computer vision related to domain adaptation and open set recognition (see Sec. 2 for more details).

In the standard framework of *Unsupervised Domain Adaptation* (DA, Saenko et al. (2010)), labeled source and unlabeled target data are drawn from two different marginal distributions that cover the same set of categories. The setting is transductive, so the target is available at training time and is used for both adaptation and evaluation. Learning solutions for this task have flourished in the last decade but they remain ineffective in the more realistic *open set domain adaptation* (OSDA, Panareda Busto and Gall (2017); Saito et al. (2018)) scenario where the source and target data contain both shared (known) and exclusive (unknown) classes. Here, forcing adaptation without recognizing the outlier samples leads to negative transfer (Rosenstein et al., 2005), adding confusion in the final known class recognition task. *Open set recognition* (Scheirer et al., 2013) and *outlier detection* focus on cases in which the source classifier also need to detect samples that belong to none of the training classes. A related line of research is that of *Positive-Unlabeled learning* (PU, Denis et al. (2005)) that deals with binary classification when the training data consists only of positive (P) and unlabeled (U) samples, where each unlabeled

*Corresponding author: Tel.: +43-665-651-767-32;
Email address: loghmani@acin.tuwien.ac.at (Mohammad Reza Loghmani)

sample could be either positive or negative. Most PU formulations deal with labeled and the unlabeled data drawn randomly from the same marginal distribution.

In this work, we highlight for the first time the relation and complementarity of DA and PU learning for OSDA, showing how it is possible to get the best of both worlds. We cast OSDA in the theoretical framework of PU learning by considering the source samples as P and the target samples as U. Our DA solution exploits PU learning to detect unknown target samples and avoid negative transfer, while extending PU learning to the case of uneven data distributions. Specifically, the main original contributions of this paper can be summarized in the following three points. **(1) We present a novel formulation of the non-negative risk estimation of PU learning** (Kiryo et al., 2017) inspired by the cross-domain robustness of reconstruction-based features (Bousmalis et al., 2016; Ghifary et al., 2016). We introduce the Positive-Unlabeled Reconstruction Encoding (PURE) algorithm that trains an autoencoder to reconstruct the known samples and map the unknown samples to a semantically void vector (Sec. 4). **(2) We integrate domain adversarial learning**, extending PURE to OSDA (OSDA-PURE), while avoiding negative transfer (Sec. 5). **(3) Finally, we propose a new evaluation metric for OSDA that penalizes large gaps in the recognition performance of known and unknown classes**. An extensive experimental analysis on the basis of our metric shows the effectiveness of OSDA-PURE with respect to its competitors (Sec. 6).

2. Related Work

Domain Adaptation. Closing the domain gap between source and target data is essential to avoid a dramatic drop in performance even for powerful deep learning methods. Discrepancy-based methods evaluate the domain shift across feature distributions with different metrics and then minimize them during training (Long et al., 2015; Le et al., 2019; Chen et al., 2018; Hu et al., 2016) . Adversarial approaches are arguably the most used ones and exploit a domain discriminator with inverted objective to promote domain confusion (Ganin et al., 2016; Russo et al., 2018). Finally, fully self-supervised tasks have recently shown to be powerful guides in learning robust cross-domain representations (Bousmalis et al., 2016; Ghifary et al., 2016; Xu et al., 2019). Still all the mentioned works deal with the closed world condition where the domains share the same set of classes. For the most challenging *open set* framework, the literature is quite limited. Panareda Busto and Gall (2017) introduced the tasks with different unknown classes in source and target. Both Saito et al. (2018) and Liu et al. (2019) focused on having only known samples in the source, while the target contains both known and unknown data. In the first work a classifier is trained to obtain a large boundary between source and target samples whereas the feature generator is trained to make the target samples far from the boundary. The second work separates the samples of known and unknown classes while weighting their importance on feature distribution alignment.

Open Set Recognition. How to make a learning model tolerant to unknown classes not seen in the training phase was a topic largely investigated in shallow learning with tailored variants of SVM and nearest-neighbor models (Scheirer et al., 2013; Mendes Júnior et al., 2017). The first deep open set approach was introduced in (Bendale and Boult, 2016), and following extensions were mainly based on generative solutions to augment the training data with synthetic unknown samples (Ge et al., 2017). A very recent work combines classification and reconstruction of input data (Yoshihashi et al., 2019). Despite their effectiveness, these methods have not been challenged with data collected across multiple domains.

Outlier Detection. Samples of unknown categories are often indicated as outliers, novel or anomalous instances, and need to be detected. Several extensions of the one-class SVM were developed for this task (Manevitz and Yousef, 2002). More recent reconstruction-based methods assume that a model optimized for inlier (positive) data will yield poorer reconstruction quality when presented with outlier (negative) data. (Xia et al., 2015), (An and Cho, 2015), and (Zhou and Paffenroth, 2017) formulate autoencoder-based approaches for this task. (Schlegl et al., 2017) used generative adversarial networks (GANs) to learn a manifold of the normal data and produce an anomaly score. (Deecke et al., 2019) instead exploited GANs and their performance in producing samples similar to an unseen test instance to score it as outlier. Note that, with respect to the more general open set case, outlier

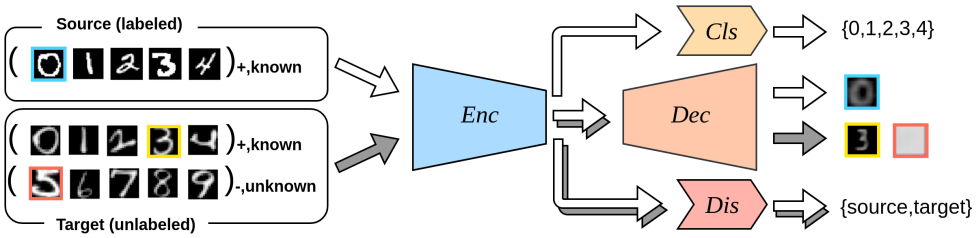

Figure 1: Schematic overview of our OSDA-PURE composed of an encoder *Enc*, a decoder *Dec*, a classifier *Cls* and a domain discriminator *Dis*. Note that the decoder reconstruction has a different effect on known and unknown target samples.

detectors do not discriminate known classes: they consider only on the binary known-unknown problem.

Positive-Unlabeled Learning. For PU, unlabeled positive and negative data are transductively available together with the annotated positive samples at training time (Hido et al., 2008; Nguyen et al., 2011). This learning framework is mainly used in outlier detection, but also in retrieval to find samples in an unlabeled data set similar to user-provided ones (Onoda et al., 2005). PU application ranges from medical to business and security with a large impact on big data for the reduction of labeling efforts (Jaskie and Spanias, 2019). The most recent publications discuss theoretical extensions then coded in shallow learning approaches (Gong et al., 2019; Kwon et al., 2019), while the only deep approach integrates the PU logic in a GAN architecture (Hou et al., 2018).

As clear, PU differs from the standard semi-supervised learning where the annotated samples for all classes are available, and at the same time it is closely related to the open set domain adaptation setting in its most simple binary form. However, labeled and the unlabeled data are generally drawn randomly from the same marginal distribution (*SCAR: Selected Completely At Random* (Elkan and Noto, 2008; du Plessis et al., 2015)). We dedicate the next section to a more rigorous discussion of PU learning basics, as useful introduction for our contribution.

3. Preliminary Background

Problem Setting. Let us consider a binary classification problem where our sample $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d$ belongs to one of the two classes with labels $y \in \{-1, +1\}$. We define as $p(\boldsymbol{x}, y)$ the underlying joint probability distribution and we indicate with $p(\boldsymbol{x})$ the marginal density. The respective class conditionals are $p_p(\boldsymbol{x}) = p(\boldsymbol{x}|y = +1)$ and $p_n(\boldsymbol{x}) = p(\boldsymbol{x}|y = -1)$, while $\pi_p = p(y = +1)$ and $\pi_n = p(y = -1) = 1 - \pi_p$ are the positive and negative class-prior probabilities.

In the standard setting for Positive-Negative (PN) learning, the data of the two classes are sampled independently from the respective marginals as $\mathcal{X}_p = \{\boldsymbol{x}_i^p\}_{i=1}^{N_p} \sim p_p(\boldsymbol{x})$ and $\mathcal{X}_n = \{\boldsymbol{x}_j^n\}_{j=1}^{N_n} \sim p_n(\boldsymbol{x})$ and the goal is to search for the optimal decision function f through *empirical risk minimization*. More precisely, if $f : \mathbb{R}^d \to \mathbb{R}$ is our decision function and we indicate with $l : \mathbb{R} \times \{\pm 1\} \to \mathbb{R}$ the *loss function* that measures with $l(t, y)$ the error incurred by predicting the output t when the ground truth is y, then we can define the *risk* of f as

$$R(f) = \mathbb{E}_{p(\boldsymbol{x},y)}[l(f(\boldsymbol{x}), y)] = \pi_p R_p^+(f) + \pi_n R_n^-(f)\,, \qquad (1)$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator, $R_p^+(f) = \mathbb{E}_{\boldsymbol{x}\sim p_p}[l(f(\boldsymbol{x}), +1)]$, and $R_n^-(f) = \mathbb{E}_{\boldsymbol{x}\sim p_n}[l(f(\boldsymbol{x}), -1)]$. The risk is empirically approximated by

$$\mathcal{L}(f) = \pi_p \widehat{R}_p^+(f) + \pi_n \widehat{R}_n^-(f)\,, \qquad (2)$$

where $\widehat{R}_p^+(f) = (1/N_p) \sum_{i=1}^{N_p} l(f(\boldsymbol{x}_i^p), +1)$ and $\widehat{R}_n^-(f) = (1/N_n) \sum_{j=1}^{N_n} l(f(\boldsymbol{x}_j^n), -1)$. Finally, using θ to parametrize the function f, the final classifier is obtained by solving $\min_\theta \mathcal{L}(f_\theta)$.

PU Learning. In the PU learning setting, the goal is to learn the binary classifier f only from positive and unlabeled data, where each unlabeled sample could be either positive or negative. Specifically, we consider the *case-control* scenario (Elkan and Noto (2008)) where two sets of data are sampled independently as $\mathcal{X}_p = \{\boldsymbol{x}_i^p\}_{i=1}^{N_p} \sim p_p(\boldsymbol{x})$ and $\mathcal{X}_u = \{\boldsymbol{x}_j^u\}_{j=1}^{N_u} \sim p(\boldsymbol{x})$. Since $\mathcal{X}_n$ is unavailable, we need a new way to approximate $R_n^-(f)$ and estimate Eq. (1). du Plessis et al. (2015) showed that starting

from $\pi_n p_n(\boldsymbol{x}) = p(\boldsymbol{x}) - \pi_p p_p(\boldsymbol{x})$, we obtain $\pi_n R_n^-(f) = R_u^-(f) - \pi_p R_p^-(f)$ where $R_u^-(f) = \mathbb{E}_{\boldsymbol{x}\sim p}[l(f(\boldsymbol{x}), -1)]$, and $R_p^-(f) = \mathbb{E}_{\boldsymbol{x}\sim p_p}[l(f(\boldsymbol{x}), -1)]$. $R(f)$ can be approximated by

$$\mathcal{L}_{PU}(f) = \pi_p \widehat{R}_p^+(f) + \widehat{R}_u^-(f) - \pi_p \widehat{R}_p^-(f)\,, \tag{3}$$

where $\widehat{R}_p^-(f) = (1/N_p)\sum_{i=1}^{N_p} l(f(\boldsymbol{x}_i^p), -1)$ and $\widehat{R}_u^-(f) = (1/N_u)\sum_{j=1}^{N_u} l(f(\boldsymbol{x}_j^u), -1)$.

Non-negative PU Learning (nnPU). Since by definition $R(f) \geq 0 \;\; \forall f$, it should also hold that $\pi_n R_n^-(f) = R_u^-(f) - \pi_p R_p^-(f) \geq 0$. However, for the empirical estimate it might not be true that $\widehat{R}_u^-(f) - \pi_p \widehat{R}_p^-(f) \geq 0$, which can cause major overfitting problems when using flexible models, such as deep neural networks, to define f. A solution to this issue is presented by Kiryo et al. (2017) through the introduction of a *non-negative risk estimator* for PU learning:

$$\mathcal{L}_{nnPU}(f) = \pi_p \widehat{R}_p^+(f) + \max\left\{0, \widehat{R}_u^-(f) - \pi_p \widehat{R}_p^-(f)\right\}. \tag{4}$$

It is worth noting that both Eq. (3) and (4) assume that the positive class-prior π_p is known. For the case-control scenario, strategies have been proposed to estimate π_p (*e.g.* du Plessis et al. (2016)). In this paper, we do not aim at obtaining a precise estimate of the class prior and simply set $\pi_p = 0.5$ throughout the experiments. The only exceptions are the experiments with selective bias (Fig. 2), where we report results with different P/N ratio in the unlabeled data and set π_p to match this ratio.

4. Autoencoder-Based Classification Loss and nnPU Risk

When the decision function f is modelled with a deep neural network, Eq. (4) is often instantiated with logarithmic loss as

$$\begin{aligned}\mathcal{L}_{LOGnnPU}(f) =& -\frac{\pi_p}{N_p}\sum_{i=1}^{N_p}\log(f(\boldsymbol{x}_i^p)) + \\ &+\max\left\{0, -\frac{1}{N_u}\sum_{j=1}^{N_u}\log(1-f(\boldsymbol{x}_j^u)) + \frac{\pi_p}{N_p}\sum_{i=1}^{N_p}\log(1-f(\boldsymbol{x}_i^p))\right\}.\end{aligned} \tag{5}$$

However, this choice produces unreliable predictions when the positive and the unlabeled samples belong to different domains. To alleviate this drawback, we need an alternative discriminative loss that is also domain agnostic. With this aim, we propose to instantiate f as an *autoencoder* (AE), a neural network architecture composed of two parts: an encoder $Enc : \mathbb{R}^d \to \mathbb{R}^e$ that projects the input into the encoding space and a decoder $Dec : \mathbb{R}^e \to \mathbb{R}^d$ that re-projects the encoded data into the input space. The energy-based learning literature indicates that the AE can be used for discriminative purposes (Zhao et al., 2017; LeCun et al., 2006). In fact, an energy-based discriminator attributes low energy (low reconstruction error) to the regions near the data manifold and high energy (high reconstruction error) to other regions. This makes an AE particularly suitable for the PU setting where there is no direct supervision on the negative data, as evidenced by Merdivan et al. (2017). In addition, the DA literature shows how the self-supervised nature of AEs makes the learned representations resilient to the difference in data domains (Bousmalis et al., 2016; Ghifary et al., 2016).

We train the AE to correctly reconstruct the positive samples while mapping the negative samples to a semantically void vector. Formally, we define the loss of $\boldsymbol{x}$ belonging to the positive and negative class as

$$l(f(\boldsymbol{x}), +1) = |\boldsymbol{x} - Dec(Enc(\boldsymbol{x}))|\,, \tag{6}$$

$$l(f(\boldsymbol{x}), -1) = |\bar{\boldsymbol{x}} - Dec(Enc(\boldsymbol{x}))|\,, \tag{7}$$

where $|.|$ denotes the absolute value function and $\bar{\boldsymbol{x}} = \kappa\mathbf{1}$ is a uniform reference vector obtained by the product of a constant κ and a d-dimensional vector of ones $\mathbf{1}$. In practice, we found a good choice to set κ to the maximum value that the input can assume. For instance, in an image classification task, $\bar{\boldsymbol{x}}$ is defined as a white image. We refer to Eq. (6) and (7) respectively as positive and negative reconstruction losses and we use them to instantiate Eq. (4) as

$$\begin{aligned}\mathcal{L}_{AEnnPU}(f) &= \pi_p \widehat{R}_p^+(f) + \max\left\{0, \widehat{R}_u^-(f) - \pi_p \widehat{R}_p^-(f)\right\} \\ &= \frac{\pi_p}{N_p}\sum_{i=1}^{N_p}|\boldsymbol{x}_i^p - Dec(Enc(\boldsymbol{x}_i^p))| + \\ &+\max\left\{0, \frac{1}{N_u}\sum_{j=1}^{N_u}|\bar{\boldsymbol{x}} - Dec(Enc(\boldsymbol{x}_j^u))| + -\frac{\pi_p}{N_p}\sum_{i=1}^{N_p}|\bar{\boldsymbol{x}} - Dec(Enc(\boldsymbol{x}_i^p))|\right\}.\end{aligned} \tag{8}$$

In the following we refer to the method minimizing this risk as **Positive and Unlabeled Reconstruction Encoding (PURE)**. At inference time, the classification output is determined with $y = sign(\tau - |\boldsymbol{x} - Dec(Enc(\boldsymbol{x}))|)$, where

τ is a threshold we set. Since the goal of minimizing the loss function in Eq. (8) is to reconstruct the unlabeled positive samples and map the unlabeled negative samples to void vectors, the absolute error $|\boldsymbol{x} - Dec(Enc(\boldsymbol{x}))|$ should be lower for positive samples and higher for negative samples. Therefore, similarly to Kato et al. (2019), we choose τ such that the top-π_p test samples with lower absolute error are classified as positive and vice versa. In Sec. 6, we provide experimental evidence to validate PURE.

5. Open-Set Domain Adaptation as a PU problem

In the OSDA setting we have annotated samples $\{(\boldsymbol{x}_i^s, c_i^s)\}_{i=1}^{N_s}$ drawn from the source domain with marginal density $p_s(\boldsymbol{x})$ and unlabeled samples $\{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$ from the target domain with marginal density $p_t(\boldsymbol{x})$. The source domain is associated with a set of *known* classes $c^s \in \{1, \dots, |C_s|\}$ that are shared with the target domain $C_s \subset C_t$, but the target covers also a set $C_{t\setminus s}$ of additional classes, which are considered *unknown*. As in closed set domain adaptation, it holds that $p_s \neq p_t$ and we further have that $p_s \neq p_t^{C_s}$ where $p_t^{C_s}$ denotes the distribution of the target domain belonging to the shared label space C_s. Ultimately, the goal of OSDA algorithms is to learn a model using the annotation of the source data to assign the target samples to either one of the $|C_s|$ shared classes or to the *unknown* class.

If we consider all the known classes as positive and the unknown classes as negative, we end up in a PU learning setting where the source data are the P set and the target data are the U set. However, since source and target belong to different domains, the SCAR assumption is not valid here. In addition, we are interested in further differentiating between the $|C_s|$ known classes. In order to tackle these problems, we equip PURE with a multi-class classifier and a domain discriminator (see Fig. 1). While PURE provides a suitable starting point for learning domain-invariant features, the domain adversarial discriminator allows the explicit minimization of the distance between the source and target domain. More formally, we extend the architecture of PURE with two new branches starting from the encoder output: a discriminator *Dis* that is trained to solve the binary source vs target problem, whose gradient backpropagates with flipped sign as in (Ganin et al., 2016) to encourage domain alignment, and a classifier *Cls* that is trained on the source samples to recognize the $|C_s|$ known classes. The final objective function of OSDA-PURE is

$$\mathcal{L} = \alpha \mathcal{L}_{AEnnPU} - \beta \mathcal{L}_{Dis} + \gamma \mathcal{L}_{Cls} \,, \tag{9}$$

where α, β, and γ are hyper-parameters that weigh each loss term in the overall objective, and the losses are

$$\begin{aligned}
\mathcal{L}_{AEnnPU} =& \frac{\pi_s}{N_s} \sum_{i=1}^{N_s} |\boldsymbol{x}_i^s - Dec(Enc(\boldsymbol{x}_i^s))| + \\
&+ \max\left\{0, \frac{1}{N_t} \sum_{j=1}^{N_t} |\bar{\boldsymbol{x}} - Dec(Enc(\boldsymbol{x}_j^t))| + - \frac{\pi_s}{N_s} \sum_{i=1}^{N_s} |\bar{\boldsymbol{x}} - Dec(Enc(\boldsymbol{x}_i^s))|\right\} , \\
\mathcal{L}_{Dis} =& - \frac{1}{N_s} \sum_{i=1}^{N_s} \log(Dis(Enc(\boldsymbol{x}_i^s))) - \frac{1}{N_t} \sum_{j=1}^{N_t} \log(1 - Dis(Enc(\boldsymbol{x}_j^t))) \,, \\
\mathcal{L}_{Cls} =& - \frac{1}{N_s} \sum_{i=1}^{N_s} c_i^s \log(Cls(Enc(\boldsymbol{x}_i^s))) \,.
\end{aligned}$$

The network is trained end-to-end in a minimax optimization scheme to converge to a saddle point of the functional of Eq. (9), using stochastic gradient descent. We use θ to indicate the network parameters and subscripts to identify the different network components, thus formally we have:

$$\begin{aligned}
(\hat{\theta}_{Enc}, \hat{\theta}_{Dec}, \hat{\theta}_{Cls}) &= \arg\min_{\theta_{Enc}, \theta_{Dec}, \theta_{Cls}} \mathcal{L}(\theta_{Enc}, \theta_{Dec}, \theta_{Cls}, \hat{\theta}_{Dis}) \\
\hat{\theta}_{Dis} &= \arg\max_{\theta_{Dis}} \mathcal{L}(\hat{\theta}_{Enc}, \hat{\theta}_{Dec}, \theta_{Dis}, \hat{\theta}_{Cls}) \,.
\end{aligned}$$

6. Experiments

6.1. Datasets

Digits. Several datasets of digit images are commonly used to study domain adaptation, namely MNIST (70k images of white digit on a black background, LeCun et al. (1998)), MNIST-M (variant of MNIST with background substituted with color photos, Ganin et al. (2016)), USPS (7k images of white digit on a black background, Friedman et al. (2001)) and SVHN (600k color images of real-world street view house numbers, Netzer et al. (2011)) with each dataset considered as a different domain. For our experiments, we always map all the samples to the highest resolution in the considered domain pair. The first 5 digits (0-4) define the positive/known/source set, and the remaining 5 digits (5-9) are unknown samples, with the unlabeled/target set covering all the 10 classes.

Cifar-Stl. Both Cifar-10 (Krizhevsky and Hinton (2009)) and Stl-10 (French et al. (2018)) are standard object classification datasets with 10 classes. Cifar-10 contains 50k training and 10k test samples, while Stl-10 has 5k training and 8k test data. For our experiments, all the images were converted to 32×32 resolution. In the open-set scenario, we define the classes *airplane, automobile, bird, cat, deer* as known and *dog, frog/monkey, horse, ship, truck* as unknown.

Office-31. This dataset (Saenko et al. (2010)) provides three domains, namely Amazon (A), DSLR (D) and Webcam (W), containing images of objects from 31 categories. Amazon contains 2820 product images from the vendor website. DSLR (534 images) and Webcam (795 images) contain similar pictures of objects taken in an office environment, with Webcam having lower quality images than DSLR. We adopt the standard open set protocol (Saito et al. (2018); Panareda Busto and Gall (2017)) where, in alphabetical order, the first 10 classes (1-10) are shared classes, and the last 10 classes (21-31) are unknowns in the target domain.

6.2. Open Set Metrics

The usual metrics adopted to evaluate OSDA are the average class accuracy over the known classes OS^*, and the accuracy of the unknown class UNK. They are generally combined to define $OS = \frac{|C_s|}{|C_s|+1} \times OS^* + \frac{1}{|C_s|+1} \times UNK$ as a measure of the overall performance. However, we argue that treating the unknown as an additional class does not provide an appropriate metric. As an example, let us consider an algorithm that is not designed to deal with unknown classes (UNK=0.0%) but has perfect accuracy over 10 known classes (OS^*=100.0%). Although this algorithm is not suitable for open set scenarios, it presents a high score of OS=90.9%. With increasing number of known classes, this effect becomes even more acute, making the role of UNK negligible. For this reason, we propose a new metric defined as the harmonic mean of OS^* and UNK, $HOS = 2\frac{OS^* \times UNK}{OS^* + UNK}$. Differently from OS, HOS provides a high score only if the algorithm performs well both on known and on unknown samples, independently of $|C_s|$. Moreover, using a harmonic mean instead of a simple average penalizes large gaps between OS^* and UNK. For a concrete example, let us consider the case where the the model classifies all samples as unknown ($OS^* = 0.0\%$, $UNK = 100.0\%$). In this case, the model converged to a trivial solution which is not informative and not suitable for open set problems. Still the mean between OS^* and UNK would be 50.0%. On the other hand HOS is 0.0%: this measure provides a clear lower bound on the evaluation metric for OSDA settings.

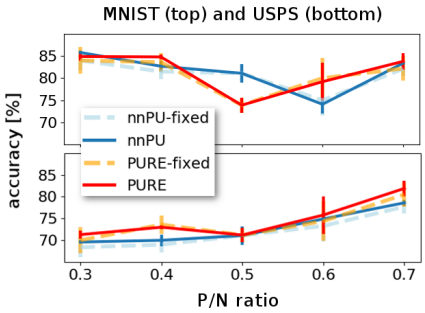

Figure 2: Mean accuracy and standard deviation over three runs in the PU setting with selection bias. The results show the performance of PURE and nnPU (Kato et al. (2019)) at different P/N ratios with (-fixed) and without fixing the prior π_p.

6.3. Results

In the first part of our experimental analysis we consider only the binary PU setting. We analyze the reconstruction loss of PURE (Eq. (8)) against the standard instantiation of nnPU with logarithmic loss (Eq. (5)), simply indicated in the following as nnPU. Moreover, we challenge nnPU and PURE with different domain shifts between the positive and the unlabeled data. In the second part, we focus on the multi-class OSDA scenario and evaluate the performance of OSDA-PURE. The hyper-parameters are selected by following Saito et al. (2018) and Liu et al. (2019) through cross-validation. We focused on learning rate and batch size in the intervals $\{10^{-2}, 10^{-4}\}$ and $\{8, 128\}$, respectively. The loss weights α, β and γ are chosen in $\{10^{-1}, 10\}$ to obtain a balanced objective function and avoid that the effect of one term would overpower the others. All the experiments are performed in a transductive setting. The implementation details are described in the supplementary materials which also reports a qualitative analysis on the reconstructed images.

Evaluating PURE. Is PURE a valid solution for the standard PU setting? We start by comparing PURE with nnPU on MNIST. We evaluate their respective Receiver

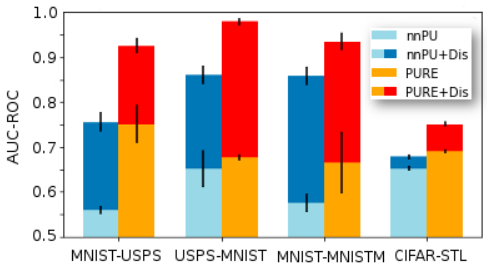

Figure 3: Mean AUC-ROC with standard deviation over three runs in the PU framework with domain shift. Results of PURE and nnPU (Kato et al. (2019)) both without and with (+ *Dis*) the domain discriminator.

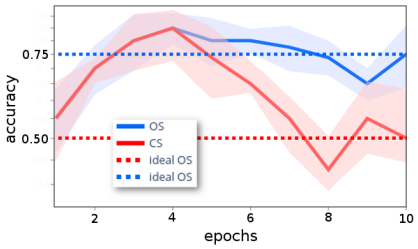

Figure 4: Accuracy of *Dis* at different learning epochs for the MNISTM-MNIST experiment in the closed set (CS) and open set (OS) scenario.

Operating Characteristic (ROC) curve when varying the sensitivity of the unknown detector by modifying the threshold τ to go from zero to complete recall. The area under the ROC curve (AUC-ROC) is similar for nnPU (0.995±0.001) and PURE (0.996±0.001) showing that the reconstruction-based non-negative risk estimator of PURE is meaningful and reliable.

Are nnPU and PURE resilient to selection bias? A mild domain shift between positive labeled and unlabeled data can be due to selection bias with the P set containing *easier* positive samples than the U set. We reproduce the setting recently studied in (Kato et al., 2019) and we compare the accuracy of nnPU with that of PURE when considering different P/N ratios in the U data. In this setting, we assume that the true value of π_p, which coincides with the considered P/N ratio, is known. For both nnPU and PURE, we choose τ such that the top-π_p test samples are classified as positive. The solid lines in Fig. 2 show that, in presence of selection bias, both nnPU and PURE provide results firmly above chance with PURE slightly outperforming nnPU in the USPS case. It is noteworthy that both methods perform well even when the P/N ratio is skewed (*e.g.* , P/N=0.3 and P/N=0.7). To test the robustness to the estimation of π_p, we repeat the experiments by setting $\pi_p = \tau = 0.5$ for all P/N ratios. The dashed lines in Fig.2 show that both nnPU and PURE maintain comparable accuracies to the case where the true prior is used.

Are nnPU and PURE resilient to cross-dataset domain shift? We investigate the challenging case where P and U belong to different domains by testing for classification across datasets (*e.g.* MNIST-USPS means that P is from MNIST and U is from USPS). Fig. 3 shows the AUC-ROC of both methods on four different domain shifts. In this setting, where the distance between the P and U distributions is larger than in the selection bias case, nnPU shows all its limits, with a performance close to chance (AUC-ROC=0.5) for MNIST-USPS and MNISTM-MNIST. PURE outperforms nnPU in all cases, with an advantage of up to +0.19 in the MNIST-USPS case. We also investigate the effects of adding a domain discriminator *Dis* to nnPU and PURE. Fig. 3 shows that both methods greatly benefit from *Dis*. Still, PURE+*Dis* steadily outperforms nnPU+*Dis* in all considered cases. These results clearly indicate that both the AE and the adversarial domain discriminator independently contribute to the domain-invariance of the learned features.

Evaluating OSDA-PURE. We compare our method against three baselines: OSVM, DANN+O and MMD+O. *OSVM* trains an Open Set SVM (Jain et al., 2014) on features generated by a CNN pre-trained on ImageNet. *MMD+O* adapts the closed set DA method of Long et al. (2015), based on minimizing the Maximum Mean Discrepancy, to the OSDA scenario by adding an OSVM. Similarly, *DANN+O* adapts the closed set DA method DANN (Ganin et al., 2016), based on domain adversarial training, to the OSDA scenario by adding an OSVM. A more detailed description of these baselines is provided in (Saito et al., 2018). Naturally, we also benchmark against the state-of-the-art methods[1]: OSBP (Saito et al., 2018) and STA (Liu et al., 2019).

How does OSDA-PURE perform on standard benchmark datasets? Following (Saito et al., 2018), we test our OSDA-PURE both on digits recognition and on ob-

[1] Baktashmotlagh et al. (2019) report the *OS* metric, not *OS**, and no public code is available. This prevents a fair comparison since it is not possible to disentangle the contribution of the known and unknown class on the results.

Table 1: Average class accuracy for known classes (OS^*), unknown classes (*UNK*), and both known and unknown classes measured with the *OS* and *HOS* metrics in the open set domain adaptation scenario for digits classification. The **top** and ***second best*** results are highlighted with bold fonts.

DIGITS																				
Method	MNISTM-MNIST				SVHN-MNIST				USPS-MNIST				MNIST-USPS				AVG			
	OS	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*
OSVM	60.8	62.3	61.0	61.5	54.3	63.1	10.5	18.0	43.1	32.3	***97.5***	48.5	79.8	77.9	***89.0***	83.1	59.1	57.7	65.7	61.4
MMD+O	46.5	47.1	46.6	46.8	55.9	64.7	12.2	20.5	62.8	58.9	82.1	68.6	80.0	79.8	81.0	80.4	68.0	68.8	58.4	63.2
DANN+O	56.8	58.4	57.0	57.6	62.9	***75.3***	0.70	1.4	84.4	***92.4***	0.9	1.8	33.8	40.5	44.3	42.3	60.4	69.4	15.3	25.1
OSBP	***91.5***	**94.7**	***75.5***	***84.0***	***63.0***	59.1	***82.5***	***68.9***	***92.3***	91.2	**97.8**	***94.4***	***92.1***	**94.9**	78.1	85.7	***84.7***	85.0	***83.5***	***83.2***
STA	72.3	85.8	5.5	10.3	**76.9**	**75.4**	**84.4**	**79.6**	92.2	91.3	96.7	93.9	**93.0**	**94.9**	83.5	***88.8***	83.6	**86.9**	67.5	68.2
OSDA-PURE	**92.5**	***93.9***	**85.5**	**89.5**	61.9	59.8	72.4	65.5	**97.2**	**97.2**	97.2	**97.2**	91.6	***92.0***	**89.3**	**90.6**	**85.8**	***85.7***	**86.1**	**85.7**

OBJECT CLASSIFICATION																				
Method	OFFICE-31 A-D				OFFICE-31 A-W				CIFAR-STL				STL-CIFAR				AVG			
	OS	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*	*OS*	*OS**	*UNK*	*HOS*
OSVM	59.6	59.1	64.6	61.7	57.1	55.0	78.1	64.5	46.7	45.3	53.5	49.1	24.1	19.3	48.4	27.6	46.9	44.7	61.2	50.7
MMD+O	47.8	44.3	***82.8***	57.7	41.5	36.2	***94.5***	52.3	45.2	43.5	53.6	48.0	24.9	20.5	46.5	28.5	39.9	36.1	69.4	47.3
DANN+O	40.8	35.6	**92.8**	51.5	31.0	24.3	**98.0**	38.9	44.5	43.2	50.7	46.7	30.3	26.4	49.4	34.5	36.7	32.5	***72.7***	42.9
OSBP	76.6	***76.4***	78.6	**77.5**	74.9	74.3	80.9	***77.5***	36.1	27.6	**78.8**	40.9	21.2	6.2	**96.1**	11.6	52.2	46.1	**83.6**	***51.9***
STA	**76.7**	**81.3**	30.7	44.6	**80.7**	**87.4**	13.7	23.3	***66.2***	***63.8***	***78.1***	**70.2**	**55.0**	**52.7**	***66.3***	**58.7**	**69.7**	**71.3**	47.2	49.3
OSDA-PURE	68.9	70.0	57.9	63.4	***80.3***	***80.8***	75.3	**78.0**	**69.9**	**68.6**	72.4	**70.4**	***52.4***	***51.9***	54.8	***53.3***	***67.9***	***67.8***	65.1	**66.8**
OSDA-PURE + init	***74.0***	75.0	64.0	***69.1***	79.7	80.4	72.7	76.4	-	-	-	-	-	-	-	-	-	-	-	-

ject classification. For all experiments, we report *OS*, OS^*, *UNK*, and *HOS*, focusing on this last metric as a measure of overall performance. The top part of Table 1 presents the results on the digits datasets, also including the MNISTM-MNIST case that was not considered in (Saito et al., 2018) and for which we ran both OSBP and STA by using the code provided by the authors. OSDA-PURE outperforms the competing methods in three out of four tasks and presents the highest average *HOS*. In the SVHN-MNIST case, STA firmly outperforms all other methods. However, STA achieves poor results on MNISTM-MNIST, highlighting an instability in the performance that is present also in the object classification tasks. The bottom part of Table 1 presents the results of object classification on Office-31 and the (CIFAR, STL) pair. Office-31 has some well-known issue: unbalanced class statistics, noisy labels and very few samples per domain (Cicek and Soatto (2019); Venkateswara et al. (2017)). Since it is a landmark dataset for DA, we still provide experiments on this dataset, but focusing only on the A-D and A-W pairs because A in the only domain with at least $1k$ samples. All the reported results on Office-31 are obtained using as backbone network AlexNet pre-trained on ImageNet (Krizhevsky et al. (2012)). For OSDA-PURE, training from scratch a decoder that is a mirrored version of AlexNet ($\sim$ 60 million parameters) would not be feasible from the limited amount of samples of this dataset. Thus we defined a *truncated autoencoder* with *Dec* composed of a single fully connected (fc) layer that is trained to reconstruct the input to the last fc layer of *Enc*. By following (Ghifary et al., 2016), we also re-ran the OSDA-PURE experiments initializing the decoder with the transpose weights of the corresponding encoder layers (OSDA-PURE + init). The results on the object classification cases show that OSDA-PURE is the only method that maintains a good performance across all tasks. In fact, OSBP performs well on A-D and A-W and poorly on CIFAR-STL and STL-CIFAR, while STA presents the opposite behavior. It is worth having a closer look at the performance of STA in the Office-31 cases. If we focus only on *OS*, STA shows the best results. However, this metric masks its poor performance on unknown samples that is well represented by *HOS*. Finally, it is interesting to compare the performance of OSDA-PURE with DANN+O since our method can be thought of as an extension of DANN with an additional decoder. Due to the effects of negative transfer, DANN+O shows even lower performance than OSVM in seven out

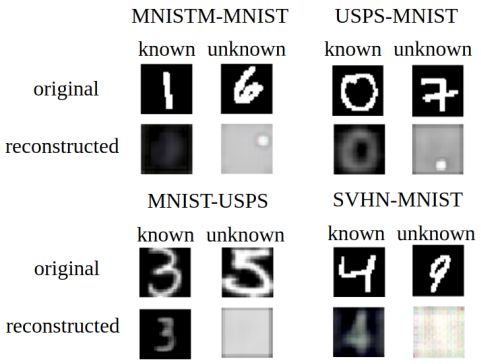

Figure 5: Qualitative analysis on the reconstructed target images from the PURE-OSDA digits experiments. Samples from known classes tend to be reconstructed to keep the original label, while samples from the unknown classes map to a mostly uniform image.

of eight cases. On the contrary, OSDA-PURE avoids negative transfer and largely outperforms both OSVM and DANN+O thanks to the contribution of *Dec* in identifying the unknown classes.

Is there an internal equilibrium between PURE and the domain discriminator? During the learning process, our PU reconstruction loss moves apart target features from source features, while the domain discriminator aligns source and target distributions. These two antagonistic forces actually collaborate to isolate the unknown target samples while reducing the domain shift among the shared classes. Fig. 4 shows the accuracy of *Dis* on the target samples for MNISTM-MNIST. The desired condition in the closed set (CS) scenario is complete confusion across domains with accuracy 0.5, while in the OS scenario half of the classes (shared by source and target) should be confused (acc. 0.5) and half should be perfectly recognized as belonging to the target (accuracy 1.0) with an expected overall accuracy of $(0.5 \times 0.5 + 0.5 \times 1) = 0.75$. After an initial phase needed by *Enc* to learn and produce domain invariant features, the performance of *Dis* converges to the expected values. Note that *Dis* does not have any explicit information on the label difference between the two domains and reaches this performance through the adversarial game with the PURE loss. This confirms that the training objective of Eq. (9) is meaningful.

Is it possible to explain/interpret the result of the method? AEs have the side advantage of visual transparency which allows us to explore the inner working of OSDA-PURE also through a qualitative data visualization. Fig. 5 shows the reconstruction effect on the target samples belonging to known and unknown classes in the digits experiments. We can see how the network is actually able to distinguish between the two cases, mapping the samples either to a meaningful digit or to an almost uniform image. Recall that we are not interested in a high-quality reconstruction and the reconstruction error is only used as a proxy for the known/unknown classification. As an example, the MNISTM-MNIST shift in Fig. 5 shows that the reconstruction of the known sample is very blurry and mostly black, but is clearly different from the reconstruction of the unknown sample, thus allowing an accurate known/unknown prediction. Another interesting insight comes from the known examples of the MNIST-USPS and the SVHN-MNIST shift in the figure: instead of replicating the input, the AE encodes the input into prototypical examples in the feature space which is then decoded to an image with a recognizable digit.

7. Discussion and Conclusions

In this work, we propose a novel method to tackle the challenging problem of open set domain adaptation by casting it into the theoretical framework of PU learning. Our OSDA-PURE gets the best of both worlds: (a) it removes the SCAR assumption in PU learning by exploiting the self-supervised power of AEs and domain adversarial training, and (b) it isolates the unknown target samples reducing the effect of negative transfer through a novel reconstruction-based PU risk estimator. Experiments in the PU learning setting show that our AE-based risk estimator is clearly superior to the standard logarithmic instantiation when the P and U sets belong to different domains. Experiments in the OSDA setting show that OSDA-PURE: (i) is the only method that consistently improves over the OSVM-based baselines; (ii) outperforms all competitors in six out of nine cases; (iii) has the highest average performance in both the digit recognition cases and the object classification cases, by a large margin in the latter. Sample reconstruction may be difficult in case of data scarcity, but recent work have shown that other self-supervised tasks are suitable for cross-domain generalization (Carlucci et al. (2019); Gidaris et al. (2018)). We plan to investigate this direction for future PU-based

OSDA approaches. Overall, our contribution sheds new light on both the field of PU and open set domain adaptation, and we believe it can be a starting point to further boost research in both areas.

Acknowledgement

M. R. L. was funded by the European Union's H2020 program under the Marie Skłodowska-Curie grant agreement No. 676157, project ACROSSING.

References

An, J., Cho, S., 2015. Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE 2.

Baktashmotlagh, M., Faraki, M., Drummond, T., Salzmann, M., 2019. Learning factorized representations for open-set domain adaptation, in: ICLR.

Bendale, A., Boult, T., 2016. Towards open set deep networks, in: CVPR.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D., 2016. Domain Separation Networks, in: NeurIPS.

Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T., 2019. Domain generalization by solving jigsaw puzzles, in: CVPR.

Chen, Y., Yang, C., Zhang, Y., 2018. Deep domain similarity adaptation networks for across domain classification. Pat. Rec. Let. (PRL) 112, 270–6.

Cicek, S., Soatto, S., 2019. Unsupervised domain adaptation via regularized conditional alignment, in: ICCV.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., Kloft, M., 2019. Image anomaly detection with generative adversarial networks, in: MLKDD.

Denis, F., Gilleron, R., Letouzey, F., 2005. Learning from positive and unlabeled examples. Theoretical Computer Science 348, 70–83.

Elkan, C., Noto, K., 2008. Learning classifiers from only positive and unlabeled data, in: ACM SIGKDD.

French, G., Mackiewicz, M., Fisher, M., 2018. Self-ensembling for visual domain adaptation, in: ICLR.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer series in statistics.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. JMLR 17, 2096–2030.

Ge, Z., Demyanov, S., Garnavi, R., 2017. Generative openmax for multi-class open set classification, in: BMVC.

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D., Li, W., 2016. Deep reconstruction-classification networks for unsupervised domain adaptation, in: ECCV.

Gidaris, S., Singh, P., Komodakis, N., 2018. Unsupervised representation learning by predicting image rotations, in: ICLR.

Gong, C., Shi, H., Liu, T., Zhang, C., Yang, J., Tao, D., 2019. Loss decomposition and centroid estimation for positive and unlabeled learning. IEEE T-PAMI .

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., Kanamori, T., 2008. Inlier-based outlier detection via direct density ratio estimation, in: ICDM.

Hou, M., Chaib-Draa, B., Li, C., Zhao, Q., 2018. Generative adversarial positive-unlabeled learning, in: IJCAI.

Hu, J., Lu, J., Tan, Y., Zhou, J., 2016. Deep transfer metric learning. IEEE Transactions on Image Processing 25, 5576–5588.

Jain, L.P., Scheirer, W.J., Boult, T.E., 2014. Multi-class open set recognition using probability of inclusion, in: ECCV.

Jaskie, K., Spanias, A., 2019. Positive and unlabeled learning algorithms and applications: A survey, in: IISA, pp. 1–8.

Kato, M., Teshima, T., Honda, J., 2019. Learning from positive and unlabeled data with a selection bias, in: ICLR.

Kiryo, R., Niu, G., du Plessis, M.C., Sugiyama, M., 2017. Positive-unlabeled learning with non-negative risk estimator, in: NeurIPS.

Krizhevsky, A., Hinton, G., 2009. Learning multiple layers of features from tiny images. Master's thesis, Dep. Computer Science, Univ. of Toronto .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: NeurIPS.

Kwon, Y., Kim, W., Sugiyama, M., Paik, M.C., 2019. Principled analytic classifier for positive-unlabeled learning via weighted integral probability metric. Machine Learning .

Le, T.N., Habrard, A., Sebban, M., 2019. Deep multi-wasserstein unsupervised domain adaptation. Pat. Rec. Let. (PRL) 125, 249–55.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F.J., 2006. A tutorial on energy-based learning, in: Predicting Structured Data, MIT Press.

Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q., 2019. Separate to adapt: Open set domain adaptation via progressive separation, in: CVPR.

Long, M., Cao, Y., Wang, J., Jordan, M.I., 2015. Learning transferable features with deep adaptation networks, in: ICML.

Manevitz, L.M., Yousef, M., 2002. One-class svms for document classification. J. Mach. Learn. Res. 2, 139–154.

Mendes Júnior, P.R., de Souza, R.M., Werneck, R.d.O., Stein, B.V., Pazinato, D.V., de Almeida, W.R., Penatti, O.A.B., Torres, R.d.S., Rocha, A., 2017. Nearest neighbors distance ratio open-set classifier. Machine Learning 106, 359–386.

Merdivan, E., Loghmani, M.R., Geist, M., 2017. Reconstruct & crush network, in: NeurIPS.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., 2011. Reading digits in natural images with unsupervised feature learning, in: NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Nguyen, M.N., Li, X.L., Ng, S.K., 2011. Positive unlabeled learning for time series classification, in: IJCAI.

Onoda, T., Murata, H., Yamada, S., 2005. One class support vector machine based non-relevance feedback document retrieval, in: IJCNN.

Panareda Busto, P., Gall, J., 2017. Open set domain adaptation, in: ICCV.

du Plessis, M.C., Niu, G., Sugiyama, M., 2015. Convex formulation for learning from positive and unlabeled data, in: ICML.

du Plessis, M.C., Niu, G., Sugiyama, M., 2016. Class-prior estimation for learning from positive and unlabeled data, in: ACML.

Rosenstein, M.T., Marx, Z., Kaelbling, L.P., Dietterich, T.G., 2005. To transfer or not to transfer, in: NeurIPS Workshop on Transfer Learning.

Russo, P., Carlucci, F.M., Tommasi, T., Caputo, B., 2018. From source to target and back: symmetric bi-directional adaptive gan, in: CVPR.

Saenko, K., Kulis, B., Fritz, M., Darrell, T., 2010. Adapting visual category models to new domains, in: ECCV.

Saito, K., Yamamoto, S., Ushiku, Y., Harada, T., 2018. Open set domain adaptation by backpropagation, in: ECCV.

Scheirer, W., Rocha, A., Sapkota, A., Boult, T., 2013. Toward open set recognition. IEEE T-PAMI 35, 1757–1772.

Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G., 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, in: IPMI.

Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S., 2017. Deep hashing network for unsupervised domain adaptation, in: CVPR.

Xia, Y., Cao, X., Wen, F., Hua, G., Sun, J., 2015. Learning discriminative reconstructions for unsupervised outlier removal, in: ICCV.

Xu, J., Xiao, L., López, A.M., 2019. Self-supervised domain adaptation for computer vision tasks. IEEE Access 7, 156694–156706.

Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., Naemura, T., 2019. Classification-reconstruction learning for open-set recognition, in: CVPR.

Zhao, J., Mathieu, M., LeCun, Y., 2017. Energy-based generative adversarial networks, in: ICLR.

Zhou, C., Paffenroth, R.C., 2017. Anomaly detection with robust deep autoencoders, in: ACM SIGKDD.