Open-set Domain Adaptation via Joint Error based Multi-class Positive and Unlabeled Learning

Dexuan Zhang[1], Thomas Westfechtel[1], and Tatsuya Harada[1,2]

[1] The University of Tokyo
[2] RIKEN

Abstract. Open-set domain adaptation aims to improve the generalization performance of a learning algorithm on a more realistic problem of open-set domain shift where the target data contains an additional unknown class that is not present in the source data. Most existing algorithms include two phases that can be described as closed-set domain adaptation given heuristic unknown class separation. Therefore, the generalization error cannot be strictly bounded due to the gap between the true distribution and samples inferred from heuristics. In this paper, we propose an end-to-end algorithm that tightly bound the risk of the entire target task by positive-unlabeled (PU) learning theory and the joint error from domain adaptation. Extensive experiments on various data sets demonstrate the effectiveness and efficiency of our proposed algorithm over open-set domain adaptation baselines.

1 Introduction

Given enough annotated training data, deep learning models can significantly improve the performance across a wide variety of machine-learning tasks but usually cannot generalize well to new domains, as it is commonly assumed that the training and test data are drawn from the same distribution. In practice, however, this assumption can be violated by several factors, such as the change in light, noise, the angle at which the image is captured, and different types of sensors, which is referred to as the domain shift that can harm the performance when predicting the test data. As a solution, domain adaptation (DA) aims to transfer the knowledge learned from a source domain, which is typically fully labeled, into a different (although related) target domain.

As a more realistic setting [28], open-set domain adaptation allows the target data to contain an additional unknown category, covering all irrelevant classes not present in the source domain. The main idea of open-set unsupervised domain adaptation (OUDA) approaches [1, 4, 8, 17, 28] is to jointly learn a classifier from a hypothesis space for known and unknown classes in the source and target domains. According to [2], the target error can be bounded by the source error, the discrepancy distance between domains, the joint error coming from the conditional shift [38], and the open-set risk that indicates the precision of unknown class recognition. Open-set risk contributes substantially to the error

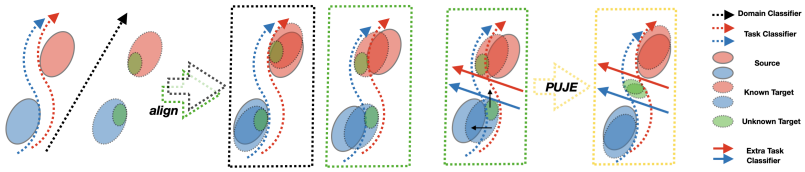
bound when a large amount of target data is from the unknown class. However, the latter two terms cannot be explicitly computed.

Unsupervised domain adaptation (UDA) methods [9, 19–21, 27, 32] show effectiveness in adapting unlabeled data to new domains by distribution alignment, but can fail to learn discriminative class boundaries, especially when the domain shift is large. Most existing methods tend to ignore the joint error and only focus on minimizing the discrepancy between domains, where samples from different classes can be grouped if the domain shift is large enough. In that case, the joint error becomes non-negligible, and the target error cannot be strictly bounded [9, 38]. [37] provided a solution to this problem by incorporating the joint error into the target error upper bound in the unsupervised setting, and we generalize this idea towards the open-set domain adaptation setting.

Despite the promising performance, existing OUDA approaches [1, 3, 4, 8, 15, 17, 28, 34] lack an essential theoretical analysis of the generalization error for the target risk. Some works claim to derive a rigorous target error bound, while the minimization of joint error and the open-set risk are not theoretically guaranteed [7, 22]. Moreover, most existing algorithms can be described as closed-set UDA after separation of the unknown [3, 22, 34, 35]. In that case, the open-set risk cannot be controlled since the generalization error is also affected by the gap between empirical known target distribution and inferred known target samples, which becomes hard to analyze.

Fig. 1a illustrates the potential problems of existing methods. Generally, it is impossible to perfectly distinguish the unknown such that the following marginal distribution alignment for UDA part is conducted under a large label shit. In that case, the joint error would be crucial as there exists a trade-off between marginal discrepancy and joint error [38]. Even if somehow we manage to obtain a perfect separation, some target data class will be dragged towards the outside of the corresponding decision boundary. In addition, the unknown class cannot be grouped as a single cluster. Fig. 1b intuitively explains the core concept of our algorithm PUJE Sec. 3, where we introduce two additional task classifiers and minimize the predictive discrepancy between the classifiers with same color, such that the unknown class will be pushed away from source data and grouped as a single cluster. Furthermore, those target data outside the corresponding decision boundary will be aligned back to the correct places.

OUDA problems can be considered as multi-class positive-unlabeled learning problems (MPU) [36] with domain shift, where all the shared classes appearing in both source and target domains are positive and the unknown class is negative. Therefore, to minimize the aforementioned joint error and open-set risk to achieve a tighter error bound for open-set adaptation, instead of unknown separation and closed-set UDA, we combine PU learning theory and joint error-based target error bound into an end-to-end learning framework. The following methods also introduced PU learning to tackle OUDA problems. [18] required the unsupervised source-like reconstruction for target data via encoder-decoder models whose performance cannot be guaranteed especially for large domain shift and complex dataset; [10] only dealt with the open-set label shift and assumed an identical



(a) existing methods **(b)** proposed method

Fig. 1: Intuitive explanation of the difference between our algorithm PUJE and existing methods. (a) existing methods do not explicitly minimize joint error and cannot group unknown class as a single cluster; (b) our proposal is an upper bound of joint error which can address large domain shift and group unknown class into a single cluster.

class conditional distribution of each instance across domains; [35] tackled a slightly different task where the target data is changing over time and it did not cover joint error and the gap between the empirical and inferred known target distribution. Compared with previous works, the main contributions of this paper can be summarized as follows:

- We propose an end-to-end algorithm with theoretical analysis for open-set problem setting by PU learning-induced joint error based target error bound;
- We define a novel discrepancy measurement, namely open-set margin discrepancy (OMD), which allows us to smoothly apply the strategy in PU learning to the target error bound in domain adaptation;
- Extensive experiments confirm the effectiveness and efficiency of the proposal.

2 Preliminaries

In this section, we introduce notations, problem settings, and theoretical definitions for the tasks of closed-set unsupervised domain adaptation, open-set unsupervised domain adaptation, and multi-class positive and unlabeled learning.

Definition 1 (Closed-set Unsupervised Domain Adaptation). *For a classification task, the learning algorithm has access to a set of n labeled points $\hat{S} = \{(x_s^i, y_s^i) \in (\mathcal{X} \subseteq \mathbb{R}^D \times \mathcal{Y} = \{1, ..., K\})\}_{i=1}^n$ from the source domain S, and a set of m unlabeled points $\hat{T} = \{(x_t^i) \in \mathcal{X}\}_{i=1}^m$ from a different target domain T. With training samples drawn i.i.d from both domains, the goal is to learn an optimal target classifier $f : \mathcal{X} \to \mathcal{Y}$.*

Definition 2 (Multi-class Positive and Unlabeled Learning). *For a classification task, the learning algorithm has access to a set of n labeled points $\hat{P} = \{(x_p^i, y_p^i) \in (\mathcal{X} \times \mathcal{Y}' = \{1, ..., K-1\})\}_{i=1}^n$ from the positive domain P, and a set of m unlabeled points $\hat{U} = \{(x_u^i) \in \mathcal{X}\}_{i=1}^m$ from the unlabeled domain U, where $y_u^i \in \mathcal{Y}$. With training samples drawn i.i.d from positive and unlabeled domains, the goal is to learn an optimal classifier $f : \mathcal{X} \to \mathcal{Y}$ for the unlabeled domain. Here, U includes samples from unknown class K, but the class conditional distributions for known classes remain invariant between P, U.*

Definition 3 (Open-set Unsupervised Domain Adaptation). *For a classification task, the learning algorithm has access a set of n labeled points* $\hat{S}' = \{(x^i_{s'}, y^i_{s'}) \in (\mathcal{X} \times \mathcal{Y}')\}^n_{i=1}$ *from the incomplete source domain* S'*, and a set of m unlabeled points* $\hat{T} = \{(x^i_t) \in \mathcal{X}\}^m_{i=1}$ *from a different target domain* T*, where* $y^i_t \in \mathcal{Y}$*. With training samples drawn i.i.d from both domains, the goal is to learn an optimal target classifier* $f : \mathcal{X} \to \mathcal{Y}$*.*

We show the relation between open-set domain adaptation and multi-class positive and unlabeled learning as follows, which includes the hits on how to apply the domain adaptation algorithm in the open-set scenario.

- In Definition 1, there is a domain shift between the distributions of the samples from the source and target domains such that a classifier trained on S cannot generalize on T. A typical solution is to align the feature distributions of both domains by a feature extractor [9, 32].
- In Definition 2, the conditional distributions of the samples from known class are identical for positive and unlabeled domains. As proved in [36], the expected error on the unlabeled domain can be approximated by the expected error on the positive domain and the probability that the unlabeled sample has not been classified as unknown.
- In Definition 3, the incomplete source domain S' can be regarded as the positive domain P in MPU. The target domain T can be treated as U with a domain shift. Assume that we can bridge the feature distributions with UDA techniques, it becomes possible to address open-set problems with well-established positive and unlabeled (PU) learning theories.

3 PU Learning induced Joint Error based OUDA (PUJE)

In this section, inspired by [37], we first rigorously reformulate the joint error-based target upper bound to facilitate the further derivation of a practical objective function for OUDA. To estimate the expectation over the complete source domain S, we deploy the theory from multi-class positive and unlabeled learning to address the open-set problems where the source domain S' is incomplete. We lastly show the entire loss of PUJE by combining the aforementioned joint error and PU learning.

Theorem 1 (Approximated Joint Error based Target Upper Bound[3]). *Given the output space* $\mathcal{K} = \{k | k \in \mathbb{R}^K : \sum_{y \in \mathcal{Y}} k[y] = 1, k[y] \in [0, 1]\}$*, let* $f_S, f_T : \mathcal{X} \to \mathcal{K}$ *be the true labeling functions for the source and target domains respectively, whose outputs are one-hot vectors denoting the corresponding labels of inputs. Let* $\epsilon : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ *denote a distance metric and* $\epsilon_D(f, f') := \mathbb{E}_{x \sim D} \epsilon(f(x), f'(x))$ *measure the expected disagreement between the outputs of* $f, f' : \mathcal{X} \to \mathcal{K}$ *over a distribution* D *on* $\mathcal{X}$*. For* $\forall f^*_S \in \mathcal{H}_S \subseteq \mathcal{H}, \forall f^*_T \in \mathcal{H}_T \subseteq \mathcal{H}, \forall h \in \mathcal{H} : \mathcal{X} \to \mathcal{K}$ *where* $h(x)[y]$ *indicates the probability of* $x \in \mathcal{X}$ *labeled as* $y \in \mathcal{Y}$*, the expected target error is bounded by*

[3] see proofs in the supplemental material

$$\epsilon_T(h) \leq \epsilon_S(h) + \epsilon_T(f_S^*, f_T^*) + \epsilon_T(h, f_S^*) - |\epsilon_S(f_S^*, f_T^*) - \epsilon_S(h, f_T^*)| + \theta \quad (1)$$

$$\theta = 2\epsilon_T(f_S, f_S^*) + \epsilon_S(f_S, f_S^*) + 2\epsilon_S(f_T^*, f_T) + \epsilon_T(f_T^*, f_T) = \theta_{f_S} + \theta_{f_T} \quad (2)$$

According to Theorem 1, we show that the expected target error $\epsilon_T(h) := \epsilon_T(h, f_T)$ is bounded by the expected source error $\epsilon_S(h) := \epsilon_S(h, f_S)$, the discrepancy between domains $[\epsilon_T(f_S^*, f_T^*) + \epsilon_T(h, f_S^*) - |\epsilon_S(f_S^*, f_T^*) - \epsilon_S(h, f_T^*)|]$, and the deviation from true labeling functions θ. The following Assumption 1 allows us to derive the generalization error of Theorem 1 such that we can ignore the residual term θ and relate target error bound with the empirical estimation of source error and the discrepancy between domains.

Assumption 1. *Let $\hat{S}, \hat{T}$ denote empirical samples with finite size from source and target domains. Assume that there exists approximated labeling functions f_S^*, f_T^* such that the empirical deviation $\hat{\theta}_{f_S}, \hat{\theta}_{f_T}$ are close enough to zero.*

Definition 4 (Empirical Rademacher Complexity). *Let $\mathcal{F}$ be a class of real-valued functions $: \mathcal{X} \to \mathbb{R}$ and $\hat{D} = \{x_1, ..., x_m\}$ a finite sample drawn i.i.d. from a distribution D, the empirical Rademacher Complexity of $\mathcal{F}$ is defined by*

$$\hat{\Re}_{\hat{D}}(\mathcal{F}) = \frac{1}{m}\mathbb{E}_\sigma[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i)], \quad (3)$$

where σ_i is an independent uniform random variable taking values in $\{-1, +1\}$.

Theorem 2 (Generalization Error[3]). *Given Theorem 1 and Definition 4, Assumption 1 and function space $\mathcal{F}_{f,f'}^{\epsilon} : \mathcal{X} \to \epsilon(f(X), f'(X))$ bounded by $M > 0$, for any $\delta > 0$, with probability at least $1 - 2\delta$, for $\forall h \in \mathcal{H}$:*

$$\begin{aligned}\epsilon_T(h) \leq \epsilon_{\hat{S}}(h) + \epsilon_{\hat{T}}(f_S^*, f_T^*) + \epsilon_{\hat{T}}(h, f_S^*) - |\epsilon_{\hat{S}}(f_S^*, f_T^*) - \epsilon_{\hat{S}}(h, f_T^*)| + 2[\hat{\Re}_{\hat{S}}(\mathcal{F}_{h,f_S}^{\epsilon}) \\ + \hat{\Re}_{\hat{T}}(\mathcal{F}_{f_S^*,f_T^*}^{\epsilon}) + \hat{\Re}_{\hat{T}}(\mathcal{F}_{h,f_S^*}^{\epsilon}) + \hat{\Re}_{\hat{S}}(\mathcal{F}_{f_S^*,f_T^*}^{\epsilon}) + \hat{\Re}_{\hat{S}}(\mathcal{F}_{h,f_T^*}^{\epsilon}) + \hat{\Re}_{\hat{S}}(\mathcal{F}_{f_S,f_S^*}^{\epsilon}) \quad (4) \\ + 2\hat{\Re}_{\hat{S}}(\mathcal{F}_{f_T,f_T^*}^{\epsilon}) + \hat{\Re}_{\hat{T}}(\mathcal{F}_{f_T,f_T^*}^{\epsilon}) + 2\hat{\Re}_{\hat{T}}(\mathcal{F}_{f_S,f_S^*}^{\epsilon})] + 3M(5\sqrt{\frac{\log\frac{9}{\delta}}{2m}} + 6\sqrt{\frac{\log\frac{9}{\delta}}{2n}})\end{aligned}$$

In the following part, we introduce the techniques from PU learning to estimate the expectation over source domain S by the incomplete source domain S' and target domain T for the open-set scenario.

Definition 5 (Unknown Predictive Discrepancy). *Let $\upsilon : \mathcal{K} \times \mathcal{K} \to \mathbb{R}$ denote the Unknown Predictive Discrepancy as a distance metric and $\upsilon_D(f, f') := \mathbb{E}_{x \sim D}\upsilon(f(x), f'(x))$ measure the expected disagreement between the K-th outputs of $f, f' : \mathcal{X} \to \mathcal{K}$ over a distribution D on $\mathcal{X}$. Let $e^K : \mathcal{X} \to [0, ..., 1] \in \mathcal{K}$ denote a function that can predict any input as unknown. The deviation from e^K for a hypothesis $h \in \mathcal{H}$ is referred to as the shorthand $\upsilon_D(h) := \upsilon_D(h, e^K)$ that measures the probability that samples from D have not been classified as unknown.*

Assumption 2. *Let $S^i = P_S(x|y=i), T^i = P_T(x|y=i)$ denote class conditional distributions, $S' = P_S(x|y \neq K), T' = P_T(x|y \neq K)$ indicate incomplete domains that do not contain unknown class S^K, T^K. Given a feature extractor $g : \mathcal{X} \subseteq \mathbb{R}^D \to \mathcal{Z} \subseteq \mathbb{R}^F$, assume that the feature space can be aligned by UDA techniques such that $Z^K = P_{S^K}(z) = P_{T^K}(z), Z' = P_{S'}(z) = P_{T'}(z)$.*

Proposition 1. *Let $h \in \mathcal{H}_{F_g} : \mathcal{Z} \to \mathcal{K}$ denote the decomposed hypothesis where $h \circ g \in \mathcal{H}$. Given Definition 5, Assumption 2, $\upsilon_{S^K}(h \circ g) = \upsilon_{T^K}(h \circ g)$ that represents probability that samples from S^K have not been classified as unknown class, is equivalent to $\epsilon_{S^K}(h \circ g)$ that measures the classification error on unknown class according to the definition.*

Lemma 1 (Estimated Source Error[3]). *Let $\sum_{i=1}^K \pi_S^i = 1, \sum_{i=1}^K \pi_T^i = 1$ denote the label distribution of S and T respectively. Given Proposition 1, the expected error on S can be estimated by the error on S' and Unknown Predictive Discrepancy (Definition 5) on T with a mild condition that $\pi_S^K = \pi_T^K = 1 - \alpha$:*

$$\epsilon_S(h \circ g) = \alpha[\epsilon_{S'}(h \circ g) - \upsilon_{S'}(h \circ g)] + \upsilon_T(h \circ g) \tag{5}$$

Remark 1. According to Definition 5, minimizing $\upsilon_T(h \circ g)$ means classifying target samples as the unknown class. In practice, a multiplier $\beta < 1$ is applied on $\upsilon_T(h \circ g)$ to prevent all target samples from being recognized as unknown.

To reformulate the intractable discrepancy $\epsilon_S(f_S^*, f_T^*), \epsilon_S(h, f_T^*)$ in Theorem 1, we decompose the approximated labeling functions given feature extractor g such that $f_S^\star \circ g = f_S^*, f_T^\star \circ g = f_T^*$ where $f_S^\star \in \mathcal{H}_{S_g} \subseteq \mathcal{H}_{F_g}, f_T^\star \in \mathcal{H}_{T_g} \subseteq \mathcal{H}_{F_g}$. We further define ϵ in a way such that, under some circumstances, it is equivalent to the Unknown Predictive Discrepancy for $\forall f \in \mathcal{H}_{F_g}$[3]:

$$\epsilon_{T^K}(f \circ g, f_T^\star \circ g) = \upsilon_{T^k}(f \circ g, f_T^\star \circ g) \tag{6}$$

Corollary 1. *Given Assumption 2, the feature extractor g can align the feature distributions of source and target domain such that the expectation over S^K is equal to that over T^K, which leads Eq. (6) to: $\epsilon_{S^K}(f \circ g, f_T^\star \circ g) = \epsilon_{T^K}(f \circ g, f_T^\star \circ g) = \upsilon_{T^K}(f \circ g, f_T^\star \circ g) = \upsilon_{S^K}(f \circ g, f_T^\star \circ g)$.*

Lemma 2 (Estimated Discrepancy[3]). *Given Corollary 1, analogous to Lemma 1, the discrepancy measured on source domain $\epsilon_S(f_S^*, f_T^*), \epsilon_S(h, f_T^*)$ can be estimated by the discrepancy on S' and Unknown Predictive Discrepancy (Definition 5) with a mild condition that $\pi_S^K = \pi_T^K = 1 - \alpha$:*

$$\begin{cases} \epsilon_S(f_S^\star \circ g, f_T^\star \circ g) & = \alpha[\epsilon_{S'}(f_S^\star \circ g, f_T^\star \circ g) - \upsilon_{S'}(f_S^\star \circ g, f_T^\star \circ g)] + \upsilon_T(f_S^\star \circ g, f_T^\star \circ g) \\ \epsilon_S(h \circ g, f_T^\star \circ g) & = \alpha[\epsilon_{S'}(h \circ g, f_T^\star \circ g) - \upsilon_{S'}(h \circ g, f_T^\star \circ g)] + \upsilon_T(h \circ g, f_T^\star \circ g) \end{cases} \tag{7}$$

Definition 6 (Overall Loss of PUJE). *Let $\hat{S}', \hat{T}$ denote empirical data from S', T. Given a feature extractor $g : \mathcal{X} \subseteq R^D \to \mathcal{Z} \subseteq \mathbb{R}^F$, decomposed hypothesis and approximated labeling functions $h, f_S^\star, f_T^\star \in \mathcal{H}_{F_g} : \mathcal{Z} \to \mathcal{K}$, replacing the intractable terms in Theorem 2 with Eqs. (5) and (7) and ignoring the Rademacher Complexity terms, the overall loss of PUJE can be regrouped as the sum of classification loss L_{cls} and the discrepancy between domains L_{dis}:*

$$\begin{aligned} L_{puje} = {} & \alpha[\epsilon_{\hat{S}'}(h \circ g) - \upsilon_{\hat{S}'}(h \circ g)] + \beta\upsilon_{\hat{T}}(h \circ g) + \epsilon_{\hat{T}}(f_S^\star \circ g, f_T^\star \circ g) + \epsilon_{\hat{T}}(h \circ g, f_S^\star \circ g) \\ & -|\upsilon_{\hat{T}}(f_S^\star \circ g, f_T^\star \circ g) - \upsilon_{\hat{T}}(h \circ g, f_T^\star \circ g) + \alpha[\epsilon_{\hat{S}'}(f_S^\star \circ g, f_T^\star \circ g) \\ & +\upsilon_{\hat{S}'}(h \circ g, f_T^\star \circ g) - \epsilon_{\hat{S}'}(h \circ g, f_T^\star \circ g) - \upsilon_{\hat{S}'}(f_S^\star \circ g, f_T^\star \circ g)]| \\ = {} & L_{cls}(h; g) + L_{dis}(f_S^\star, f_T^\star, h; g) \end{aligned} \tag{8}$$

4 Methodology

In this section, we first define Approximated Labeling Function Space $\mathcal{H}_{S_g}, \mathcal{H}_{T_g}$ such that the intractable $L_{dis}(f_S^\star, f_T^\star, h; g)$ due to unknown $f_S^\star, f_T^\star$ can be upper bounded by $\sup_{f'_S \in \mathcal{H}_{S_g}, f'_T \in \mathcal{H}_{T_g}} L_{dis}(f'_S, f'_T, h; g)$. To fulfill Eq. (6), we propose Open-set Margin Discrepancy to quantify ϵ that measures the disagreement between classifiers. We lastly show the entire training algorithm.

4.1 Approximated Labeling Function Space

Proposition 2. *Let $\mathcal{H}_S, \mathcal{H}_T : \mathcal{X} \to \mathcal{K}$ be two sets of functions that can minimize a part of the empirical residual $\hat{\theta}_S, \hat{\theta}_T$ respectively, $f_S^* \in \mathcal{H}_S, f_T^* \in \mathcal{H}_T$ must hold as f_S^*, f_T^* can minimize the entire $\hat{\theta}$ given Assumption 1. Accordingly, decomposed functions $f_S^\star, f_T^\star$ must lie in $\mathcal{H}_{S_g} = \{f | \forall f \circ g \in \mathcal{H}_S\}, \mathcal{H}_{T_g} = \{f | \forall f \circ g \in \mathcal{H}_T\}$ such that a sufficient condition of the following inequality: $L_{dis}(f_S^\star, f_T^\star, h; g) \leq \sup_{f'_S \in \mathcal{H}_{S_g}, f'_T \in \mathcal{H}_{T_g}} L_{dis}(f'_S, f'_T, h; g)$ is fulfilled.*

Definition 7 (Approximated Labeling Function Space). *Let $L_{\mathcal{H}_S}, L_{\mathcal{H}_T}$ denote a part of the empirical residual $\hat{\theta}_{f_S}, \hat{\theta}_{f_T}$ respectively. Let $f_S^\star \circ g = f_S^* \in \mathcal{H}_S, f_T^\star \circ g = f_T^* \in \mathcal{H}_T$ denote the decomposed approximated labeling functions where $f_S^\star \in \mathcal{H}_{S_g} \subseteq \mathcal{H}_{F_g}, f_T^\star \in \mathcal{H}_{T_g} \subseteq \mathcal{H}_{F_g}$. Given Proposition 2, Approximated Labeling Function Space $\mathcal{H}_{S_g}, \mathcal{H}_{T_g}$ can be defined as the sets whose members $f'_S, f'_T \in \mathcal{H}_{F_g}$ can minimize $L_{\mathcal{H}_S}, L_{\mathcal{H}_T}$:*

$$\begin{cases} \mathcal{H}_{S_g} = \{f'_S | \arg\min_{g, f'_S \in \mathcal{H}_{F_g}} L_{\mathcal{H}_S}(f'_S; g) := L_{cls}(f'_S; g)\} \\ \mathcal{H}_{T_g} = \{f'_T | \arg\min_{g, f'_T \in \mathcal{H}_{F_g}} L_{\mathcal{H}_T}(f'_T; g) := (1-\gamma) L_{cls}(f'_T; g) + \gamma L_{reg}\} \end{cases} \tag{9}$$

$\mathcal{H}_{S_g}$ consists of functions minimizing the expected error on S, which can be estimated by Lemma 1. To build a reliable function space $\mathcal{H}_{T_g}$ without target labels, we approximate the target error with the weighted average ($\gamma \in [0, 1]$) of error rate on labeled samples and a semi-supervised regularization term $L_{reg} = L_{ent} + L_{pse} + \omega L_{con}$.

Regularized Entropy Minimization As introduced in [11,27,31], we impose a class balance prior that can penalize classifiers with complex decision boundaries on entropy minimization [12] to yield a more sensible solution:

$$\begin{aligned} L_{ent} = &-\mathbb{E}_{x\in\hat{T}} \sum\nolimits_{y\in\mathcal{Y}} f'_T(g(x))[y] \log f'_T(g(x))[y] \\ &+ \sum\nolimits_{y\in\mathcal{Y}} \mathbb{E}_{x\in\hat{T}} f'_T(g(x))[y] \log \mathbb{E}_{x\in\hat{T}} f'_T(g(x))[y] \end{aligned} \tag{10}$$

Pseudo Labeling As introduced in [30, 31], for input $x \in \hat{T}$ and its random augmentation x' [5], we minimize cross entropy for x with pseudo labels of x':

$$L_{pse} = -\mathbb{E}_{x\in\hat{T}} \log f'_T(g(x))[\arg\max\nolimits_{y\in\mathcal{Y}} h(g(x'))[y]] \tag{11}$$

Consistency Regularization As introduced in [14, 29], we penalize the difference of the outputs for input $x \in \hat{T}$ and its random augmentation x':

$$L_{con} = \mathbb{E}_{x\in\hat{T}} |f'_T(g(x)) - f'_T(g(x'))| \tag{12}$$

4.2 Open-set Margin Discrepancy

Definition 8 (Induced Labeling Function). *Let $f : \mathcal{X} \to \mathcal{K}$ denote a multi-class labeling function, where $f(x)[y]$ indicating y-th element of the output $f(x)$ for the probability of x classified as y. Thus an induced labeling function $l \circ f : \mathcal{X} \to \mathcal{Y}$ is given by: $l(f(x)) = \arg\max_{y\in\mathcal{Y}} f(x)[y]$.*

Definition 9 (Open-set Margin Discrepancy (OMD)). *Let $y = l(f(x)), y' = l(f'(x))$ denote the predictions on input x from $f, f' : \mathcal{X} \to \mathcal{K}$ given Definition 8. Open-set Margin Discrepancy between f, f' over a distribution D is given by*

$$\epsilon_D(f, f') = \mathbb{E}_{x\sim D}[\mathrm{omd}(f(x), f'(x))] \tag{13}$$

$$\begin{aligned} \mathrm{omd}(f(x), f'(x)) = \max(&|\log(1 - f(x)[y]) - \log(1 - f'(x)[y])|, \\ &|\log(1 - f(x)[y']) - \log(1 - f'(x)[y'])|) \end{aligned} \tag{14}$$

For OUDA, generally the classifier is not discriminative for unknown class K due to the lack of labeled samples to reduce the misclassification rate. To address this, we model $\log(1 - f(x)[K])$ that has a more stable gradient.

Remark 2. According to Definition 5, $\upsilon_D(f, f')$ measures the expectation of disagreement on the K-th output between two functions $f, f' : \mathcal{X} \to \mathcal{K}$ over a distribution D, which can also indicate the difference between probabilities of categorizing inputs as known classes. To fulfill Eq. (6), we quantify $\upsilon_D(f, f')$ as:

$$\upsilon_D(f, f') = \mathbb{E}_{x\sim D} |\log(1 - f(x)[K]) - \log(1 - f'(x)[K])| \tag{15}$$

Algorithm 1 PUJE

Input: incomplete source data $\hat{S}'$; unlabeled target data $\hat{T}$
Output: labeling functions f'_S, f'_T; feature extractor g; hypothesis h
Parameter: trade-off parameter λ; learning rate η; known class ratio estimator α
for *epoch* $= 1, 2, \ldots$ **do**
  Step 1: estimate known class ratio α on T with prediction of g, h
  Step 2: optimize g, f'_S, f'_T to satisfy the approximated labeling function space
    $(g, f'_S, f'_T) \leftarrow (g, f'_S, f'_T) + \eta \Delta(g, f'_S, f'_T)$
    $\Delta(g, f'_S, f'_T) = -\frac{\partial(L_{\mathcal{H}_S}(f'_S; g) + L_{\mathcal{H}_T}(f'_T; g))}{\partial(g, f'_S, f'_T)}$
  Step 3: maximize the discrepancy w.r.t. f'_S, f'_T within the function space
    $(f'_S, f'_T) \leftarrow (f'_S, f'_T) + \eta \Delta(f'_S, f'_T)$
    $\Delta(f'_S, f'_T) = -\frac{\partial(L_{\mathcal{H}_S}(f'_S; g) + L_{\mathcal{H}_T}(f'_T; g) - \lambda L_{dis}(f'_S, f'_T, h; g))}{\partial(f'_S, f'_T)}$
  Step 4: minimize the entire target error bound w.r.t. g, h for fixed f'_S, f'_T
    $(g, h) \leftarrow (g, h) + \eta \Delta(g, h)$
    $\Delta(g, h) = -\frac{\partial(L_{cls}(h; g) + \lambda L_{dis}(f'_S, f'_T, h; g))}{\partial(g, h)}$
end for

4.3 Training Algorithm

Given the feature extractor $g : \mathcal{X} \subseteq \mathbb{R}^D \to \mathcal{Z} \subseteq \mathbb{R}^F$ and hypotheses $h, f'_S, f'_T \in \mathcal{H}_{F_g} : \mathcal{Z} \to \mathcal{K}$, we introduce a trade-off parameter λ to balance the classification loss and discrepancy as Eq. (16). During each epoch, we first estimate the known class ratio α (initially set to 1) in target data by the predictions of h. Then we train the hypothesis h, approximated labeling functions f'_S, f'_T, feature extractor g to minimize the error on labeled data L_{cls} with hypothesis space constraints $L_{\mathcal{H}_S}, L_{\mathcal{H}_T}$, while optimizing the discrepancy L_{dis} adversarially by the min-max game over g and f'_S, f'_T (Algorithm 1):

$$\begin{cases} \min_{f'_S, f'_T \in \mathcal{H}_{F_g}} L_{\mathcal{H}_S}(f'_S; g) + L_{\mathcal{H}_T}(f'_T; g) - \lambda L_{dis}(f'_S, f'_T, h; g) \\ \min_{h \in \mathcal{H}_{F_g}, g} L_{\mathcal{H}_S}(f'_S; g) + L_{\mathcal{H}_T}(f'_T; g) + L_{cls}(h; g) + \lambda L_{dis}(f'_S, f'_T, h; g) \end{cases} \tag{16}$$

4.4 Intuition

In this section, we intuitively explain how our method can align the source and target domains while separating the unknown in Fig. 2. According to [24], unlike the entropy minimization that tends to cut through unlabeled data, consistency regularization helps to draw a manifold-aware decision boundary. Based on this conclusion, the decision boundary of f'_T induced by consistency regularization L_{con} can preserve the cluster structure (Fig. 2a). According to Algorithm 1, f'_S tries to maximize L_{dis} while classify target samples as unknown class to minimize $L_{\mathcal{H}_S}(f'_S; g)$ (Definition 7 and Lemma 1), which can lead to a new decision boundary shown in Fig. 2b. Meanwhile, feature extractor g tries to minimize L_{dis} by pushing target samples close to f'_S towards source clusters,

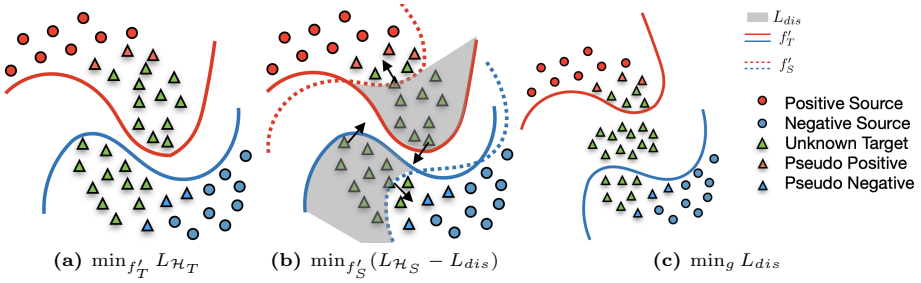


Fig. 2: Intuitive explanation of the mechanism of the proposed PUJE in Algorithm 1. (a) $L_{\mathcal{H}_T}$ includes consistency regularization L_{con} that helps to draw a manifold-aware decision boundary for f'_T (Step 2); (b) f'_S can increase L_{dis} by classifying most of the unlabeled target samples as unknown, which also decreases $L_{\mathcal{H}_S}$ (Step 3); (c) feature extractor g pushes unlabeled target samples close to the decision boundary of f'_S towards source cluster, while those close to f'_T towards outside to reduce L_{dis}, which can lead to a separated unknown cluster (Step 4).

while those close to f'_T towards outside such that unlabeled target samples far from source clusters can be eventually separated and grouped into a new cluster for the unknown progressively as illustrated in Fig. 2c.

5 Evaluation

We evaluated our proposal on two benchmarks, Office-Home and Syn2Real-O. In the implementation, the L_{reg} weight γ and the trade-off parameter λ were set to 0.1 and 0.01 according to [37]. In addition, we empirically set PU coefficient β to 0.03 and consistency coefficient ω to 20 for Office-Home and 5 for Syn2Real-O. All experiments were conducted with the ImageNet [6] pre-trained ResNet-50 [13] as the feature extractor g and 2-layer linear networks for classifiers f'_S, f'_T, h. We trained the model by Stochastic Gradient Descent optimizer with a 0.001 learning rate annealed according to [9], 24 batch-size, a momentum of 0.9. We quantitatively compare our results against various OUDA baselines, including OSBP [28], STA [17], DAOD [8], PGL [22], ROS [3], OSLPP [34] and ANNA [15].

Evaluation Metrics To evaluate the proposed method and the baselines, we utilize the widely used measures [22, 28], i.e., normalized accuracy for all classes (OS), normalized accuracy for the known classes only (OS*) and harmonic mean HOS=2(OS* $\times$ UNK)/(OS* + UNK) [3, 15, 18, 34].

Office-Home *[33]* is a widely-used domain adaptation benchmark, which consists of 15,500 images from 65 categories and four domains: Art (Ar), Clipart (Cl), Product (Pr), and Real-World (Rw). Following the same splits used in previous methods [22], we select the first 25 classes in alphabetical order as the known classes and group the rest as the unknown.

Table 1: Accuracy of ResNet-50 model fine-tuned on Office-Home dataset (OS)

METHOD	Pr→Ar			Pr→Cl			Pr→Rw			Rw→Ar			RW→Cl			Rw→Pr		
	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS
OSBP	47.1	65.3	64.6	38.3	48.7	48.3	27.0	81.6	79.5	37.1	73.5	72.1	29.3	55.3	54.3	37.7	81.9	80.2
STA	**77.0**	48.4	49.5	**95.4**	40.8	42.9	**59.1**	77.3	76.6	**71.2**	68.6	68.7	**61.0**	45.4	46.0	**58.9**	74.5	73.9
DAOD	44.3	67.7	66.8	44.7	60.3	59.7	40.8	85.0	83.3	49.8	73.2	72.3	47.4	60.4	59.9	56.8	82.8	81.8
PGL	34.7	73.7	72.2	38.4	59.2	58.4	27.6	84.8	82.6	6.1	**81.5**	78.6	25.1	66.8	65.0	38.0	84.8	83.0
PUJE	38.4	**74.8**	**73.4**	42.2	**64.4**	**63.6**	37.9	**87.7**	**85.8**	23.4	81.0	**78.8**	39.2	**71.2**	**70.0**	39.1	**88.2**	**86.3**

METHOD	Ar→Cl			Ar→Pr			Ar→Rw			Cl→Rw			Cl→Pr			Cl→Ar			MEAN		
	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS	UNK	OS*	OS
OSBP	28.6	57.2	56.1	25.8	77.8	75.8	23.0	85.4	83.0	33.0	77.2	75.5	16.7	71.3	69.2	32.1	65.9	64.6	31.3	70.1	68.6
STA	64.1	45.9	46.6	62.0	67.2	67.0	**66.2**	76.6	76.2	**57.4**	65.2	64.9	**60.2**	57.6	57.7	**72.7**	49.3	50.2	**67.1**	59.7	60.0
DAOD	**71.1**	55.5	56.1	**66.6**	69.2	69.1	63.7	79.3	78.7	54.8	78.2	77.3	54.6	70.2	69.6	55.1	62.9	62.6	54.1	70.4	69.8
PGL	19.1	63.3	61.6	32.1	78.9	77.1	40.9	87.7	85.9	5.3	**85.9**	**82.8**	24.5	73.9	72.0	33.8	70.2	68.8	27.1	75.9	74.0
PUJE	57.1	**66.6**	**66.3**	52.1	**82.8**	**81.6**	45.2	**88.9**	**87.2**	55.9	81.8	80.8	51.4	**81.5**	**80.3**	42.5	**71.0**	**69.9**	43.7	**78.3**	**77.0**

Table 2: Accuracy of ResNet-50 model fine-tuned on Syn2Real-O dataset. * indicates our re-implementation with the officially released code.

METHOD	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	UNK	OS*	OS	HOS
OSBP	73.6	57.9	58.2	65.2	67.4	29.5	84.1	47.0	67.8	5.7	**89.1**	0.5	66.6	53.8	54.8	59.5
STA	64.1	70.3	53.7	59.4	80.8	20.8	90.0	12.5	63.2	30.2	78.2	2.7	59.1	52.2	52.7	55.4
PGL	81.5	68.3	74.2	60.6	**91.9**	**45.4**	92.2	41.0	**87.9**	**67.5**	79.2	**6.4**	49.6	**66.8**	65.5	56.9
PGL*	81.3	**77.5**	66.5	71.7	90.6	35.7	**92.4**	41.2	82.6	46.9	82.3	4.9	59.6	64.5	64.1	62.0
ANNA	41.2	54.6	39.7	59.4	51.2	25.6	82.6	54.8	78.9	10.5	76.6	2.0	67.4	48.1	49.6	56.1
PUJE	**90.7**	73.8	**75.1**	**72.4**	77.8	33.8	91.2	**58.8**	79.5	46.5	84.9	5.1	**69.0**	65.8	**66.0**	**67.3**

Syn2Real-O *[26]* is a more challenging synthetic-to-real benchmark, which is constructed from the VisDA-17 [25]. The Syn2Real-O dataset significantly increases the ratio of unknown data in the target domain to 0.9 by introducing additional unknown samples. In this dataset, the source domain contains training data from the VisDA-17 as the known set, and the target domain includes the test data from the VisDA-17 (known set) plus 50k images from irrelevant categories of MSCOCO [16] dataset (unknown set).

As reported in Tabs. 1 and 2, we observe that our method PUJE consistently outperforms the state-of-the-art results, improving OS and HOS by 3.0% and 5.3% on the benchmark datasets of Office-Home and Syn2Real-O respectively. Note that our proposed approach provides significant performance gains for the more challenging dataset of Syn2Real-O that requires knowledge transfer across different modalities. In the sub-tasks with a larger domain shift, e.g., Rw→Cl and Pr→Cl in Office-Home, we can observe similar results demonstrating the strong adaptation ability of the proposed framework. PGL reported a higher OS score in Syn2Real-O, but we cannot reproduce the results with the officially released code. During the re-implementation based on their early stop strategy, we found the UNK was around 30%, which is far below the reported score. A potential reason to explain this is that maybe they did not remove the unknown category from the source data.

Table 3: Accuracy of ResNet-50 model fine-tuned on Office-Home dataset (HOS)

METHOD	Pr→Ar			Pr→Cl			Pr→Rw			Rw→Ar			RW→Cl			Rw→Pr		
	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
ROS	64.3	57.3	60.6	71.2	46.5	56.3	78.4	70.8	74.4	70.8	67.0	68.8	73.0	51.5	60.4	80.0	72.0	75.7
OSLPP	76.2	54.6	63.6	67.1	53.1	59.3	71.2	77.0	74.0	75.0	60.8	67.2	64.3	54.4	59.0	70.8	78.4	74.4
ANNA	70.3	**63.0**	**66.5**	74.8	**54.6**	**63.1**	**78.9**	74.3	76.6	**77.3**	66.1	**71.3**	73.1	59.7	65.7	**81.0**	76.4	**78.7**
PUJE	**77.8**	56.9	65.7	**76.8**	51.2	61.5	75.2	**80.1**	**77.5**	67.3	**71.1**	69.1	**77.9**	**62.4**	**69.3**	70.5	**85.8**	77.4

METHOD	Ar→Cl			Ar→Pr			Ar→Rw			Cl→Rw			Cl→Pr			Cl→Ar			MEAN		
	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS	UNK	OS*	HOS
ROS	74.1	50.6	60.1	70.3	68.4	69.3	77.2	75.8	76.5	72.2	65.3	68.6	71.6	59.8	65.2	65.5	53.6	58.9	72.4	61.6	66.2
OSLPP	67.1	55.9	61.0	73.1	72.5	72.8	69.4	80.1	74.3	73.9	67.2	66.9	73.3	61.6	66.9	79.0	49.6	60.9	71.7	63.8	67.0
ANNA	78.7	**61.4**	**69.0**	**79.9**	68.3	73.7	**79.7**	74.1	76.8	80.2	66.9	73.0	**73.6**	64.2	68.6	73.1	**58.0**	64.7	**76.7**	65.6	70.7
PUJE	**83.5**	58.5	68.8	78.9	**79.6**	**79.3**	76.3	**82.1**	**79.1**	**80.6**	**71.2**	**75.6**	72.8	**71.4**	**72.1**	**81.2**	54.7	**65.4**	76.6	**68.8**	**71.7**

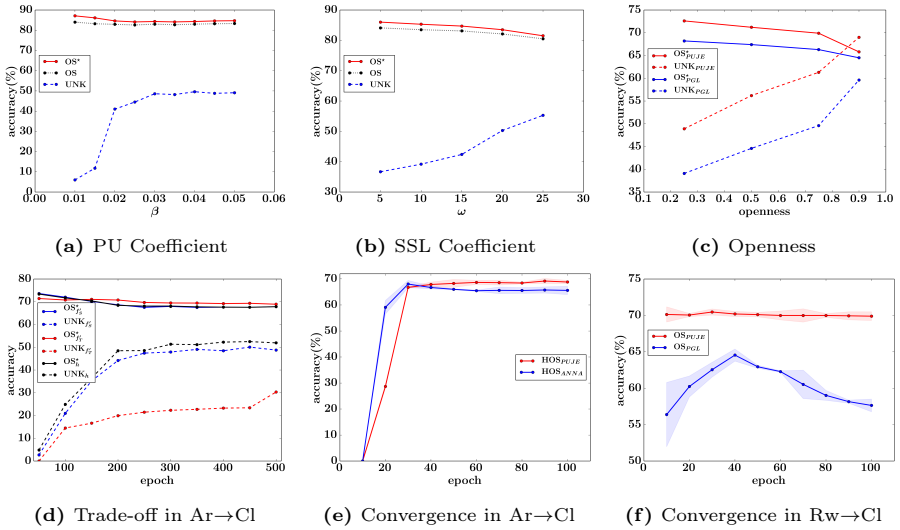
Table 4: Ablation studies of semi-supervised regularization losses. We report the Accuracy (%) on Office-Home of A → R and C → P using a ResNet50 backbone.

Config.	Ar→Pr			Cl→Rw		
	UNK	OS*	OS	UNK	OS*	OS
default	52.1	82.8	81.6	55.9	81.8	80.8
w/o L_{pse}	69.6	78.1	77.8	64.8	77.8	77.3
w/o L_{con}	26.3	81.7	79.6	33.3	80.9	79.1
w/o L_{ent}	50.2	81.1	79.9	53.9	80.5	79.5

HOS Metric Instead of OS score, [18] proposed the harmonic mean of OS* and unk, HOS=2(OS* × UNK)/(OS* + UNK) to penalize large gaps between OS* and unk. We agree that the OS score does not reflect UNK when the number of known classes grows. However, the HOS score is still unfair. OS* and UNK are equally valued in HOS, but UNK is much easier to improve than OS* which depends on the power of the adaptation algorithm. [15] utilizes a weak adaptation algorithm based on [28] that leads to a relatively loose alignment (Fig. 4) on the feature space of source and target domains such that unlabeled data is more frequently assigned to the unknown class. We believe that if a method can achieve a high OS* score, it is possible to obtain a decent HOS score by tuning hyper-parameters or additional losses. For our proposal, we can increase $\beta = 0.25$ to improve UNK at the sacrifice of OS* (Tab. 3).

Ablation Study To investigate the impact of the semi-supervised regularization, we compare three variants of the PUJE model on the Office-Home dataset shown in Tab. 4. Without L_{pse}, OS* decreases because it can no longer leverage the pseudo labels that can help to build reliable hypothesis space $\mathcal{H}_{T_g}$ (Definition 7). Without L_{con}, UNK drops as it is not likely to draw a manifold-aware decision boundary that is beneficial to the separation of unknown data (Fig. 2a).

Robustness against Varying Openness To verify the robustness of the proposed PUJE, we conducted experiments on the Syn2Real-O with the openness varying in $\{0.25, 0.5, 0.75, 0.9\}$. Here, openness is defined by the ratio of unknown



(a) PU Coefficient **(b)** SSL Coefficient **(c)** Openness

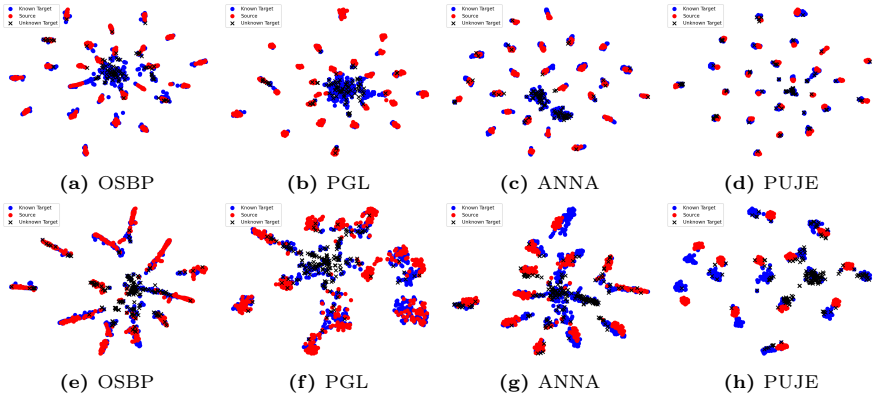
(d) Trade-off in Ar→Cl **(e)** Convergence in Ar→Cl **(f)** Convergence in Rw→Cl

Fig. 3: (a)-(b): sensitivity to varying loss coefficient β, ω verified in Ar→Pr and Cl→Rw tasks; (c): performance comparisons w.r.t. varying openness of the Syn2Real-O task; (d): training cure of the proposed algorithm that characterizes a trade-off between the classification accuracy on known classes and the unknown class; (e)-(f): convergence analysis of the Office-Home task compared to other baselines with confidence intervals.

samples in the entire target data. We show the results of PGL and PUJE in Fig. 3c. Note that the PGL approach heuristically sets the hyper-parameter according to the true unknown ratio to control the openness, while PUJE automatically estimates the weight α during the training procedure. We observe that PUJE consistently outperforms the baseline by a large margin, especially on unknown recognition, which confirms its robustness to the change in openness.

Training Curve In Fig. 3d, we illustrate the recognition performance of PUJE over training steps in the Ar→Cl task. All metrics show a performance gain over the first several steps. Then, the normalized accuracy OS* experiences a downward while the accuracy for the unknown class keeps improving, which characterizes a trade-off between OS* and UNK. In addition, the behavior of f'_S, f'_T meets the expectation of our alignment mechanism where f'_S starts to detect unknown data, which is followed by f'_T (Fig. 2b). From Figs. 3e and 3f, we observe that some previous works [15, 22] do not converge at the reported optimum, and they tackle this problem by ceasing the model updates before the score starts to decline. However, this early stop strategy can fail without enough labeled target data to validate the current performance. In contrast, our method can reach a reliable convergence.

Sensitivity to PU & ***SSL Coefficient*** We show the sensitivity of our approach to various PU and SSL loss coefficient β, ω. We tested the value of β from [0.01, 0.5]



(a) OSBP **(b)** PGL **(c)** ANNA **(d)** PUJE

(e) OSBP **(f)** PGL **(g)** ANNA **(h)** PUJE

Fig. 4: Feature space visualized by t-SNE in (a)-(d): Office-Home; (e)-(h): Syn2Real-O.

in Ar→Pr task and ω from [5, 25] in Cl→Rw task while fixing other parameters to the default setting. We can draw two observations from Figs. 3a and 3b: the OS score is relatively stable, and the unknown recognition achieves a reliable performance when coefficient β is within the interval of [0.02, 0.05]; Generally, a larger ω, which is set for lower openness, will lead to a better performance on recognizing unknown data while hurt the accuracy for known classes.

Feature Space Visualization To intuitively visualize the effectiveness of OUDA approaches, we extracted features from the baseline models (OSBP, PGL, ANNA) and our proposed model on the Ar→Cl task (Office-Home) and Syn2Real-O task with the ResNet-50 backbone [13]. The feature distributions were processed with t-SNE [23] afterward. As shown in Fig. 4, compared with baselines, our method PUJE achieves a better alignment between source and target distributions, especially when the domain shift is large. Benefiting from our joint error-based adversarial alignment mechanism, the extracted feature space, including the cluster of unknown target data, has a more discriminative class-wise decision boundary.

6 Conclusion

We have addressed the open-set domain shift problem by introducing a novel learning theory based on multi-class positive-unlabeled learning that can reduce the open-set risk and the joint error. Experiments show that our proposed learning theory performs consistently better on challenging object recognition benchmarks for open-set adaptation with significant domain gaps and label shifts.

Acknowledgements

This research is partially supported by JST Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015 and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo.

References

1. Baktashmotlagh, M., Faraki, M., Drummond, T., Salzmann, M.: Learning factorized representations for open-set domain adaptation. In: ICLR (2019)
2. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.: A theory of learning from different domains. Machine Learning **79**, 151–175 (2010)
3. Bucci, S., Loghmani, M.R., Tommasi, T.: On the effectiveness of image rotation for open set domain adaptation. In: ECCV (2020)
4. Busto, P.P., Gall, J.: Open set domain adaptation. In: ICCV (2017)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: CVPRW (2020)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition pp. 248–255 (2009)
7. Fang, Z., Lu, J., Liu, F., Xuan, J., Zhang, G.: Open set domain adaptation: Theoretical bound and algorithm. arXiv **abs/1907.08375** (2019)
8. Feng, Q., Kang, G., Fan, H., Yang, Y.: Attract or distract: Exploit the margin of open set. In: ICCV (2019)
9. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(1), 2096–2030 (2016)
10. Garg, S., Balakrishnan, S., Lipton, Z.C.: Domain adaptation under open set label shift. In: NeurIPS (2022)
11. Gomes, R., Krause, A., Perona, P.: Discriminative clustering by regularized information maximization. In: NIPS (2010)
12. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NIPS (2005)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition pp. 770–778 (2015)
14. Laine, S., Aila, T.: Temporal ensembling for semi-supervised learning. In: ICLR (2017)
15. Li, W., Liu, J., Han, B., Yuan, Y.: Adjustment and alignment for unbiased open set domain adaptation. In: CVPR (2023)
16. Lin, T.Y., Maire, M., Belongie, S.J., Bourdev, L.D., Girshick, R.B., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014)
17. Liu, H., Cao, Z., Long, M., Wang, J., Yang, Q.: Separate to adapt: Open set domain adaptation via progressive separation. In: CVPR (2019)
18. Loghmania, M.R., Vinczea, M., Tommasi, T.: Positive-unlabeled learning for open set domain adaptation. Pattern Recognition Letters **136**, 198–204 (2020)
19. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning. vol. 37, pp. 97–105. JMLR.org (2015)

20. Long, M., CAO, Z., Wang, J., Jordan, M.I.: Conditional adversarial domain adaptation. In: Advances in Neural Information Processing Systems, pp. 1640–1650. Curran Associates, Inc. (2018)
21. Long, M., Zhu, H., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70, pp. 2208–2217. JMLR.org (2017)
22. Luo, Y., Wang, Z., Huang, Z., Baktashmotlagh, M.: Progressive graph learning for open-set domain adaptation. In: ICML (2020)
23. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research **9**, 2579–2605 (2008)
24. Oliver, A., Odena, A., Raffel, C., Cubuk, E.D., Goodfellow, I.J.: Realistic evaluation of deep semi-supervised learning algorithms. In: NeurIPS (2018)
25. Peng, X., Usman, B., Kaushik, N., Hoffman, J., Wang, D., Saenko, K.: Visda: The visual domain adaptation challenge. arXiv **abs/1710.06924** (2017)
26. Peng, X., Usmana, B., Saito, K., Kaushik, N., Hoffman, J., Saenko, K.: Syn2real: A new benchmark forsynthetic-to-real visual domain adaptation. arXiv **abs/1806.09755** (2018)
27. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3723–3732 (2017)
28. Saito, K., Yamamoto, S., Ushiku, Y., Harada, T.: Open set domain adaptation by backpropagation. In: ECCV (2018)
29. Sajjadi, M., Javanmardi, M., Tasdizen, T.: Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In: NIPS (2016)
30. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In: NIPS (2020)
31. Tang, H., Chen, K.C., Jia, K.: Unsupervised domain adaptation via structurally regularized deep clustering. In: IEEE Conference on Computer Vision and Pattern Recognition (2020)
32. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. IEEE Conference on Computer Vision and Pattern Recognition pp. 2962–2971 (2017)
33. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. IEEE Conference on Computer Vision and Pattern Recognition pp. 5385–5394 (2017)
34. Wang, Q., Meng, F., Breckon, T.P.: Progressively select and reject pseudo-labelled samples for open-set domain adaptation. In: AAAI (2022)
35. Wu, J., He, J.: Domain adaptation with dynamic open-set targets. In: KDD (2022)
36. Xu, Y., Xu, C., Xu, C., Tao, D.: Multi-positive and unlabeled learning. In: IJCAI (2017)
37. Zhang, D., Westfechtel, T., Harada, T.: Unsupervised domain adaptation via minimized joint error. In: TMLR (2023)
38. Zhao, H., Combes, R.T.D., Zhang, K., Gordon, G.: On learning invariant representations for domain adaptation. In: Proceedings of Machine Learning Research. vol. 97, pp. 7523–7532. PMLR (2019)