Partial Optimal Transport Based Out-of-Distribution Detection for Open-Set Semi-Supervised Learning

Yilong Ren[1,4], **Chuanwen Feng**[2,4], **Xike Xie**[3,4], **S. Kevin Zhou**[3,4,5]
[1]School of Artificial Intelligence and Data Science, University of Science and Technology of China
[2]School of Computer Science, University of Science and Technology of China (USTC)
[3]School of Biomedical Engineering, University of Science and Technology of China (USTC)
[4]Data Darkness Lab, MIRACLE Center, Suzhou Institute for Advanced Research, USTC
[5]Key Laboratory of Precision and Intelligent Chemistry, USTC
{ylren, chuanwen}@mail.ustc.edu.cn, xkxie@ustc.edu.cn, s.kevin.zhou@gmail.com

Abstract

Semi-supervised learning (SSL) is a machine learning paradigm that utilizes both labeled and unlabeled data to enhance the performance of learning tasks. However, SSL methods operate under the assumption that the label spaces of labeled and unlabeled data are identical, which may not hold in open-world applications. In such scenarios, the unlabeled data may contain novel categories that were not presented in the labeled training data, essentially outliers. This specific challenge is referred to as the Open-set Semi-supervised Learning (OSSL) problem. In OSSL, a pivotal concern is the detection of out-of-distribution (OOD) samples within unlabeled data. Existing methods often struggle to provide effective OOD detection strategies, especially when dealing with datasets comprising a large number of training categories. In response to this challenge, we model the OOD detection problem in OSSL as a partial optimal transport (POT) problem. With POT theory, we devise a mass score function to measure the likelihood of a sample being an outlier, which enables a binary classifier for OOD detection. Further, we put forward an OOD loss, enabling the seamless integration of the binary classifier and off-the-shelf SSL methods under OSSL settings, all within an end-to-end training framework. We extensively evaluate our proposal under various datasets and OSSL configurations, consistently demonstrating the superior performance of our proposal. Codes are available at https://github.com/ryl0427/Code_for_POT_OSSL.

1 Introduction

Semi-supervised learning (SSL) is a branch of machine learning that leverages both labeled and unlabeled data to improve the performance of learning tasks [Lee and others, 2013; Laine and Aila, 2016; Tarvainen and Valpola, 2017; Berthelot *et al.*, 2019; Xie *et al.*, 2020; Sohn *et al.*, 2020]. A common assumption of SSL is that the label spaces of labeled and unlabeled data are same. However, in open-world settings, a large volume of unlabeled data need to be collected, which often contain novel categories that are unseen in the labeled data. This gives rise to a more practical setting called Open-set Semi-supervised Learning (OSSL) [Yu *et al.*, 2020], in which known categories samples, i.e., inliers, should be classified into correct classes, while unknown categories samples, i.e., outliers, should be identified. However, existing SSL approaches fall short in OSSL settings. For example, a common step in SSL is pseudo-labeling [Yang *et al.*, 2023a], which often assigns inappropriate pseudo-labels to outliers, thus posing challenges to the generalization over inliers and also the detection of outliers.

Identifying outliers, referred to as out-of-distribution (OOD) samples, stands as a pivotal challenge in the realm of OSSL. To tackle the challenge, one intuitive strategy is to integrate existing OOD detection methods into OSSL training. For example, **UASD** [Chen *et al.*, 2020] and **SAFE STUDENT** [He *et al.*, 2022] proposed to use maximum softmax prediction probability [Hendrycks and Gimpel, 2016] and energy-discrepancy score [Liu *et al.*, 2020], respectively, to detect outliers. **TOSSL** [Ma *et al.*, 2023] transformed the problem into a closed-set problem, by classifying OOD samples into a new category via distance-based methods [Lee *et al.*, 2018; Sun *et al.*, 2022]. However, these OOD detection methods are designed for the testing phase, relying on models trained over in-distribution (ID) data. In contrast, OSSL asks for OOD detection during the training phase. The latter is more challenging, because: 1) the training data comprises a mixture of ID and OOD data; 2) the training-phase detection implies working with an under-developed model. An alternative strategy relies on model predictions for OOD detection. For instance, **MTCF** [Yu *et al.*, 2020] proposed to filter ID samples using noisy label learning, and **DS3L** [Guo *et al.*, 2020] designed a weighting function for assigning weights to input samples, which is jointly trained with network parameters through bi-level optimization. It is evident that the solely dependence on neural network predictions potentially degrades the performance of OOD detection, due to its overfitting on OOD samples.

In response to these deficiencies, a current trend involves employing multiple *one-vs-all* classifiers to distinguish if unlabeled samples belong to a known class, designating those

not belonging to any specific class as OOD samples. In this context, **OpenMatch** [Saito *et al.*, 2021] introduced an open-set consistency regularization, while **T2T** [Huang *et al.*, 2021] used self-supervision and cross-modal matching, **SSB** [Fan *et al.*, 2023] utilized non-linear transformations and pseudo-negative mining, **OSP** [Wang *et al.*, 2023] proposed aliasing OOD matching and soft orthogonality regularization, to enhance *one-vs-all* classifiers. More recently, **IOMatch** [Li *et al.*, 2023] jointly considered inliers and outliers for scenarios with extremely scarce labels. Nevertheless, these methods fall short in OSSL scenarios with numerous classification categories due to the usage of K classifiers. It results in the overall error rate being computed as $1-(1-a)^K$, where a represents the error rate of each individual classifier, ultimately degrading the performance.

To this end, we introduce a novel binary OOD detector rooted in optimal transport (OT) theory, operating directly in the feature space, so that: 1) it gets rid of the reliance on the network output during the training phase, and thus avoids the overfitting deficiency in MTCF and DS3L; 2) it obviates the need for training numerous binary classifiers in one-vs-all strategies, such as T2T, OpenMatch, and IOMatch; 3) our method reallocates weights for unlabeled samples to align the unlabeled distribution with the labeled distribution, thus offering better interpretability for discerning OOD samples; 4) it improves the performance of both closed-set classification and open-set detection within OSSL.

The idea of our proposal originates in tackling the problem of distribution mismatch [Bickel *et al.*, 2009], the fundamental technical challenge of OSSL, where the labeled and unlabeled data may originate from different distributions. It is known that the OT theory [Villani, 2008; Courty *et al.*, 2014] is a mathematical tool for quantifying distribution discrepancies, providing an effective way of measuring the distribution mismatch. However, conventional OT operates under the assumption that both distributions process identical total mass, requiring complete transportation of all mass from one distribution to another. This assumption becomes problematic in the context of OSSL, where only a subset of the unlabeled data aligns with the labeled data, and the rest consists of outliers. In such cases, the assumption is untenable. Therefore, we turn to partial optimal transport (POT) [Caffarelli and McCann, 2010], a variant of OT, which allows the selective transport of a fraction of mass from one distribution to another. And yet, to the best of our knowledge, POT in OSSL remains unexplored, despite its potential utility in OSSL.

In this paper, we first formulate the outlier detection problem in OSSL as a POT problem. We seek the partial optimal transport plan between labeled and unlabeled distributions. With POT theory, we devise a mass score function (MSF) to measure the likelihood of unlabeled samples being outliers based on the mass transported between the two distributions. The MSF is calculated through the optimal transport plan of POT, where the MSF of an unlabeled sample is determined by the sum of its mass transport to the labeled distribution. To train the model in an end-to-end manner, we propose a novel OOD loss. This loss function involves predicting the input using a specially designed output layer referred to as the OOD-head, while the target value is the MSF generated by the POT process. Moreover, this approach enables us to adapt the off-the-shelf SSL methods (e.g. FixMatch) into OSSL settings via multi-task learning.

Extensive experiments across various datasets and settings consistently show the superior performance of our method in the realm of OSSL. Notably, our approach effectively addresses the performance decline witnessed in current methods when confronted with intricate scenarios featuring a substantial number of classification categories. Specifically, when applied to CIFAR-100 with 80 known classes, our method has demonstrated a remarkable improvement, achieving a noteworthy increase of over 7% in closed-set classification accuracy. The main contributions are summarized as follows.

- We directly address the distribution mismatch challenge using OT theory, modeling OOD detection in OSSL as a POT problem. This approach enhances the interpretative mechanism for identifying OOD samples.
- We introduce a novel mass score function (MSF) as a binary OOD detector. Simultaneously, our OOD detector can be integrated with off-the-shelf SSL methods through multi-task learning.
- Our method consistently surpasses current approaches, effectively addressing performance decline in complex scenarios with numerous classification categories.

2 Problem Setup

2.1 Problem Definition of OSSL

In OSSL, the labeled data consists exclusively of ID samples, while the unlabeled data and the test data include a mixture of ID and OOD samples. The goal of OSSL is to obtain a model capable of accurately classifying ID samples into their corresponding classes and distinguish OOD samples at test time.

In SSL, a widely used approach is to sample both labeled and unlabeled batches together, combining explicit supervision from labeled samples with the inherent information in unlabeled data. At each training iteration, SSL methods typically sample n labeled samples and m unlabeled samples, where each sample can be represented by a d-dimensional feature vector after undergoing the network's feature extractor. Then, the samples of labeled and unlabeled space constitute two d-dimensional point clouds, which can be represented by $\mathbf{L}=\{\mathbf{l}_i\}_{i=1}^n$ and $\mathbf{U}=\{\mathbf{u}_i\}_{i=1}^m$, respectively. We can represent the two point clouds of samples in the form of probability distributions, $\mathcal{L}$ and $\mathcal{U}$, indicating the labeled and unlabeled distributions, where δ represents the Dirac function.

$$\mathcal{L}=\sum_{i=1}^{n}\frac{1}{n}\delta_{\mathbf{l}_i}\qquad \mathcal{U}=\sum_{j=1}^{m}\frac{1}{m}\delta_{\mathbf{u}_j} \tag{1}$$

As shown in the right half of Figure 1, in OSSL, there exists a mismatch between $\mathcal{L}$ and $\mathcal{U}$. Consequently, the unlabeled distribution can be represented as the union of $\mathcal{U}_{\text{ID}}$ and $\mathcal{U}_{\text{OOD}}$, where $\mathcal{U}_{\text{ID}}$ aligns with $\mathcal{L}$, while $\mathcal{U}_{\text{OOD}}$ contrasts with it. Effectively tackling the challenges posed by OSSL necessitates addressing this inherent mismatch between $\mathcal{L}$ and $\mathcal{U}$ during the training stage.

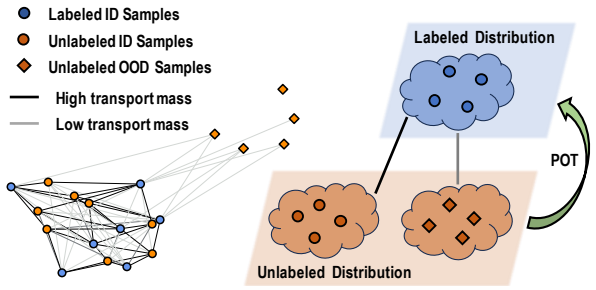

Figure 1: An illustration showcasing OOD detection within the context of OSSL through the application of partial optimal transport.

2.2 OT Preliminaries

Optimal transport (OT), as discussed in [Villani, 2021], is to seek an optimal transport plan between two measures at a minimal cost, resulting in the Wasserstein distance metric, as explained in [Figalli and Glaudo, 2021], providing a geometric way for aligning probability measures.

OT proves to be versatile in addressing distribution shift challenges. [Courty *et al.*, 2014] introduced a domain adaptation method using regularized optimal transport to align source and target domains while preserving their structures. [Feng *et al.*, 2023] introduced an OT-Filter to filter label noise under distribution shift. For open-set domain adaptation, both [Xu *et al.*, 2020] and [Yang *et al.*, 2023b] modeled with partial optimal transport. [Xu *et al.*, 2020] regulated partial optimal transport using the mean cost of transport, while [Yang *et al.*, 2023b] estimated the transport ratio. [Chapel *et al.*, 2020] proposed a PU-learning method based on partial optimal transport, and [Lu *et al.*, 2023] applied OT to the semantically coherent out-of-distribution detection task. However, as far as we know, the potential utility of OT in OSSL has not been investigated yet, despite its possible benefits in OSSL.

In the context of OSSL, a notable challenge arises from the distribution mismatch between the labeled distribution $\mathcal{L}$ and the unlabeled distribution $\mathcal{U}$. An insightful approach involves exploring the application of optimal transport theory to detect samples in $\mathcal{U}$ that are not aligned with $\mathcal{L}$, essentially $\mathcal{U}_{\text{OOD}}$.

When addressing the distribution mismatch challenge in OSSL through OT, the initial step involves defining the optimal transport problem between the labeled distribution $\mathcal{L}$ and the unlabeled distribution $\mathcal{U}$. Then, in the context of the optimal transport problem, an admissible coupling represents a joint probability distribution over two measures and must adhere to constraints on marginal distributions, ensuring complete mass transport. The set of all admissible couplings $\Pi(\mathcal{L},\mathcal{U})$ between the two distributions is given as follows.

$$\Pi(\mathcal{L},\mathcal{U}) = \left\{\mathbf{T} \in \mathbb{R}_+^{|\mathcal{L}|\times|\mathcal{U}|} | \mathbf{T}\mathbf{1}_{|\mathcal{U}|} = \mathcal{L}, \mathbf{T}^\top \mathbf{1}_{|\mathcal{L}|} = \mathcal{U}\right\} \quad (2)$$

where $\mathbf{T}$ is a coupling matrix, and T_{ij} describes the amount of mass at $\mathbf{l}_i$ transported to $\mathbf{u}_j$. Given a transport cost matrix $\mathbf{C}$, where the element C_{ij} represents the cost of transporting one unit of mass from $\mathbf{l}_i$ to $\mathbf{u}_j$, optimal transport addresses the problem of transporting $\mathcal{L}$ toward $\mathcal{U}$ with minimal cost. A coupling matrix that achieves the minimal cost is called the optimal transport plan and the minimal cost is called the optimal transport distance. The OT problem can be described as follows.

$$\min_{\mathbf{T}\in\Pi(\mathcal{L},\mathcal{U})} \langle \mathbf{C}, \mathbf{T}\rangle_F = \min_{\mathbf{T}\in\Pi(\mathcal{L},\mathcal{U})} \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}T_{ij} \quad (3)$$

However, when dealing with the obtained population-wise metric, i.e., the transport distance between $\mathcal{L}$ and $\mathcal{U}$, it is impractical to identify an individual OOD sample using OT distance. Therefore, it becomes imperative to identify OOD samples by considering the perspective of the optimal transport plan represented by $\mathbf{T}$.

2.3 POT for OSSL

Identifying OOD samples through the optimal transport plan presents a notable challenge. The conventional OT assumes that the two distributions have the same total probability mass, i.e., $\|\mathcal{L}\|_1 = \|\mathcal{U}\|_1$, and that all the mass of an unlabeled sample must be transported. Although this assumption can be well applied to the task involving identical distributions or domain adaptation, it encounters challenges in OSSL settings, where only a portion of samples in $\mathcal{U}$ aligns with $\mathcal{L}$. However, each unlabeled sample is transported to specific labeled samples, consequently disrupting the efficient transport of aligned samples from $\mathcal{U}$ to $\mathcal{L}$. At the same time, every unlabeled sample carries an identical transport mass, posing a challenge in identifying OOD samples through the transport plan.

A straightforward way is to transport the mass of ID samples between $\mathcal{U}$ and $\mathcal{L}$, while maintaining the mass of OOD samples within $\mathcal{U}$. This gives rise to the concept of POT, which allows the transport of a portion of the mass. The left half of Figure 1 offers a simulated example of OOD detection in POT, where the linewidth indicates the amount of transport mass between two samples. It conveys that the transport mass between samples of $\mathcal{L}$ and $\mathcal{U}_{\text{ID}}$ significantly outweighs that of $\mathcal{L}$ and $\mathcal{U}_{\text{OOD}}$. So, with POT, it holds good potentials for identifying individual OOD samples.

Formally, the focus of the POT problem lies in transporting a fraction $0 \le s \le min(\|\mathcal{L}\|_1, \|\mathcal{U}\|_1)$ of the mass, while simultaneously minimizing the associated cost. The set of admissible couplings, $\Pi^s(\mathcal{L},\mathcal{U})$ in this case, is defined as:

$$\Pi^s(\mathcal{L},\mathcal{U}) = \left\{\mathbf{T} \in \mathbb{R}_+^{|\mathcal{L}|\times|\mathcal{U}|} | \mathbf{T}\mathbf{1}_{|\mathcal{U}|} \le \mathcal{L}, \mathbf{T}^\top \mathbf{1}_{|\mathcal{L}|} \le \mathcal{U}, \mathbf{1}_{|\mathcal{L}|}^\top \mathbf{T}\mathbf{1}_{|\mathcal{U}|} = s\right\} \quad (4)$$

Therefore, the POT problem can be described as:

$$\min_{\mathbf{T}\in\Pi^s(\mathbf{p},\mathbf{q})} \langle \mathbf{C}, \mathbf{T}\rangle_F = \min_{\mathbf{T}\in\Pi^s(\mathbf{p},\mathbf{q})} \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}T_{ij} \quad (5)$$

Moving forward, our goal is to tackle the OSSL problem using the principles of partial optimal transport.

3 Method

3.1 Formulating POT in OSSL

In this section, our primary focus is on formalizing the OOD detection task in OSSL as a POT problem. We aim to use the

labeled distribution as a baseline distribution to detect outliers. Therefore, we redefine the labeled distribution $\mathcal{L}$ and k-fold unlabeled distribution $\mathcal{U}$ as follows.

$$\mathcal{L} = \sum_{i=1}^{n} \frac{1}{n} \delta_{\mathbf{l}_i} \qquad \mathcal{U} = \sum_{i=1}^{m} \frac{k}{m} \delta_{\mathbf{u}_i} \ (k > 1) \tag{6}$$

where $\mathbf{l}_i$ and $\mathbf{u}_i$ represent the feature of the i-th sample in labeled and k-fold unlabeled distributions, respectively. The introduced parameter k denotes the redundant mass, which enables transporting k-fold mass from each sample in $\mathcal{U}$. In this way, we represent the resulting problem as a POT problem with the following feasible set.

$$\Pi(\mathcal{L},\mathcal{U}) = \{\mathbf{T} \in \mathbb{R}_{+}^{|\mathcal{L}|\times|\mathcal{U}|} | \ \mathbf{T}\mathbf{1}_{|\mathcal{U}|} = \mathcal{L}, \mathbf{T}^{\top}\mathbf{1}_{|\mathcal{L}|} \leq \mathcal{U}\} \tag{7}$$

Incorporating equation constraints into the labeled distribution enhances the interpretability of our approach. As interpreted by Equation 7, the problem is to redistribute weights among unlabeled samples, aligning the unlabeled distribution with the labeled distribution. Moreover, the concise constraints specified in Equation 7 expedites the convergence of POT calculations.

In tackling the POT problem, the discrete optimal transport formulation, in essence, is a convex optimization problem, more precisely, a linear programming problem. Unfortunately, this linear programming problem suffers from a cubic computing complexity. To mitigate this, one way is to leverage entropic regularization, as outlined in [Cuturi, 2013], which is formulated as follows.

$$\begin{aligned} &\min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} \langle \mathbf{C}, \mathbf{T} \rangle_F - \varepsilon H(\mathbf{T}) \\ &where \quad H(\mathbf{T}) := -\sum_{i,j} T_{ij}\big(log(T_{ij}) - 1\big) \end{aligned} \tag{8}$$

where $\varepsilon > 0$ is the regularization coefficient, and $H(T)$ is the entropic regularization term. According to [Benamou *et al.*, 2015], we can recast the regularized OT problem in the language of Kullback-Leibler projections:

$$\min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} \langle \mathbf{C}, \mathbf{T} \rangle_F - \varepsilon H(\mathbf{T}) = \varepsilon \min_{\mathbf{T}\in\Pi(\mathbf{p},\mathbf{q})} KL(\mathbf{T}|e^{-\frac{\mathbf{C}}{\varepsilon}}) \tag{9}$$

The set of feasible solutions, denoted as $\mathcal{C}$ in the regularized POT context, is the intersection of two convex subspaces, i.e., $\mathcal{C}_1 \cap \mathcal{C}_2$, where

$$\mathcal{C}_1 = \big\{\mathbf{T}\mathbf{1}_{|\mathcal{U}|} = \mathcal{L}\big\} \qquad \mathcal{C}_2 = \big\{\mathbf{T}^{\top}\mathbf{1}_{|\mathcal{L}|} \leq \mathcal{U}\big\} \tag{10}$$

Obviously, the above formulation of POT for OSSL can be viewed as a convex optimization problem. Following with previous works ([Dykstra, 1983; Bauschke and Lewis, 2000]) for solving POT, we conclude our computing framework as Algorithm 1.

Algorithm 1 POT

Require: labeled distribution $\mathcal{L}$, k-fold unlabeled distribution $\mathcal{U}$, cost matrix $\mathbf{C}$, regularization ε, and error δ.
Constrains: $\mathcal{C}_1 = \big\{\mathbf{T}\mathbf{1}_{|\mathcal{U}|} = \mathcal{L}\big\} \quad \mathcal{C}_2 = \big\{\mathbf{T}^{\top}\mathbf{1}_{|\mathcal{L}|} \leq \mathcal{U}\big\}$
Initialize: $\mathbf{T}_0 = e^{-\frac{\mathbf{C}}{\varepsilon}}$, $q^{(0)} = q^{(-1)} = \mathbf{1}$, $\mathcal{C}_{n+2} = \mathcal{C}_n$
while $err < \delta$ **do**
  $\mathbf{T}^{(n)} = \underset{\mathbf{T}^{(n)}\in\mathcal{C}_n}{\arg\min} \mathbf{KL}(\mathbf{T}^{(n)}|\mathbf{T}^{(n-1)} \odot q^{(n-2)})$
  $q^{(n)} = q^{(n-2)} \odot \frac{\mathbf{T}^{(n-1)}}{\mathbf{T}^{(n)}}$
  $err = \left\|\mathbf{T}^{(n)} - \mathbf{T}^{(n-1)}\right\|$
end while
return $\mathbf{T}$

3.2 Mass Score Function: A POT Based Binary OOD Detector

In Section 3.1, we have formulated POT between probability distributions defined over labeled and unlabeled datasets. This formulation prompts the need to create a robust scoring function for assessing the probability of a sample being classified as an OOD sample. To tackle this challenge, we introduce a novel mass score function (MSF) derived from the optimal transport plan in POT. The MSF aggregates the mass transferred from unlabeled samples to labeled ones and functions as a binary OOD detector.

Next, we investigate the mechanism of MSF and explain why POT proves to be effective for OOD detection. The key idea is to leverage the discrepancy between labeled distribution $\mathcal{L}$ and the k-fold unlabeled distribution $\mathcal{U}$ to identify the portion within $\mathcal{U}$ that aligns with $\mathcal{L}$. As shown in Equation 6, we treat $\mathcal{L}$ as the reference distribution, where all samples carry identical weight (i.e., undergo the same mass transportation), assured by equality constraints. In contrast, samples in $\mathcal{U}$ may have different mass transportation, falling within the range of $[0, \frac{k}{m}]$, as guaranteed by inequality constraints.

The cost of transporting unit mass between ID samples, which includes both labeled and unlabeled samples, is relatively low. In line with the principle of POT, only a fraction of the mass within the unlabeled distribution can be feasibly transported. To minimize the overall transportation cost, ID samples in unlabeled dataset obtain a higher allocation of transportation mass. This characteristic contributes to the identification of OOD samples based on the transported mass of each individual sample. Alternatively, considering this from a different perspective, the MSF adjusts the weights for unlabeled samples to align the unlabeled distribution with the labeled distribution. Samples that do not conform to the labeled distribution, namely OOD samples, naturally receive smaller weight assignments.

Thus, in order to derive the mass score function, we initially employ the cosine distance as the transport cost between labeled and unlabeled samples. Subsequently, we use Algorithm 1 to obtain the optimal transport plan. At last, we calculate MSF and scale it to the range $[0, 1]$ for training. The procedure is detailed in Algorithm 2. Following the procedure, we can derive a binary OOD detector based on POT.

3.3 OSSL Training with Proposed OOD Detector

In Section 3.2, we present a novel OOD detection method using POT between $\mathcal{L}$ and $\mathcal{U}$. While this approach is effective during the training stage for calculating the OOD score for

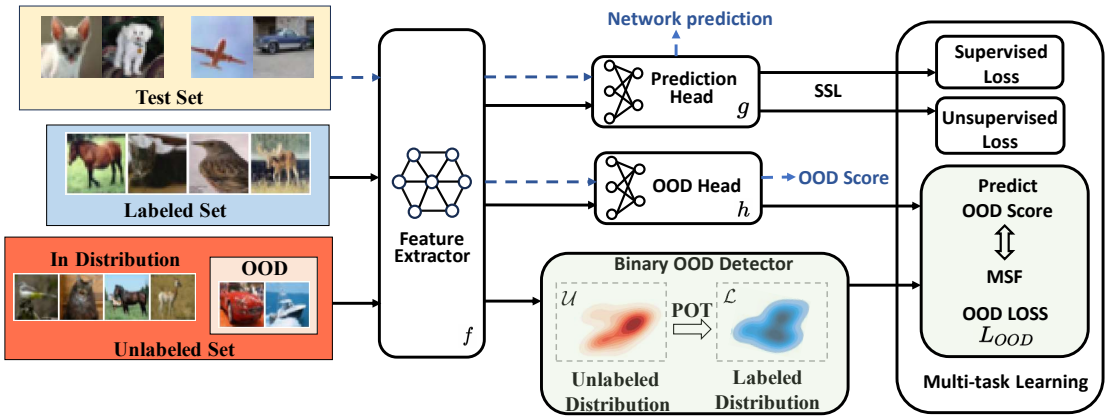

Figure 2: An illustration of our framework. Applying POT theory, we develop a Mass Score Function (MSF) in the feature space to assess the likelihood of unlabeled samples being outliers during training. For detecting OOD samples during testing, we introduce the OOD-head, which predicts the OOD score of input samples. The OOD-head is optimized through OOD loss during the training phase. Finally, we seamlessly integrate the OOD loss into off-the-shelf semi-supervised learning methods through multi-task learning.

Algorithm 2 Mass Score Function

Require: Feature of labeled and k-fold unlabeled distribution $\mathbf{L}$ and $\mathbf{U}$, entropy regularization coefficient ε, and mass redundancy parameter k.
Distribution: $\mathcal{L} = \sum_{i=1}^{n} \frac{1}{n}\delta_{\mathbf{l}_i}$ and $\mathcal{U} = \sum_{i=1}^{m} \frac{k}{m}\delta_{\mathbf{u}_i}$
Cost: calculate cosine distance matrix $\mathbf{M} = Cosine(\mathbf{L}, \mathbf{U})$
OT Plan: calculate optimal transport plan using algorithm1: $\mathbf{T} = POT(\mathcal{L}, \mathcal{U}, \mathbf{M}, \varepsilon)$
OOD Score: calculate optimal transport mass as OOD score for $\mathcal{U}$: $Score_{\mathcal{U}} = \mathbf{T}^{\top}\mathbf{1}_n$
return Scaled $Score_{\mathcal{U}}$

each unlabeled sample, challenges arise when applying POT on the test data, since there are no corresponding labeled samples available for reference. To address this problem, we investigate the OOD-head, an additional output layer, to retain the capability of detecting OOD samples, supported by POT.

In particular, we denote the score produced by the OOD-head as $Pred_{\mathcal{U}}$ and the value of MSF predicted by POT as $Score_{\mathcal{U}}$. At the training stage, the OOD-head is expected to produce scores close to the values of MSF for unlabeled samples. To this end, we resort to the following mean square error (MSE) loss.

$$L_{\text{OOD}} = \text{MSE}(Pred_{\mathcal{U}}, Score_{\mathcal{U}}) \tag{11}$$

Essentially, the OSSL task can be viewed as a combination of a semi-supervised classification task and an OOD detection task. We therefore formulate the OSSL problem as a multi-task learning problem, which is implemented by only using a simple multi-layer feedforward neural network [Caruana, 1997]. More, the two tasks share the same feature extractor and use different prediction heads. By doing this, our method can be integrated with various semi-supervised methods[1].

In addition, we outline two strategies that we have implemented to further enhance the performance of our OOD-head. Firstly, we utilize both labeled and unlabeled samples for OOD-head training, as opposed to solely relying on unlabeled samples. Leveraging supervision information from ID labeled samples, we therefore set the OOD score of each labeled sample as 1. Second, we use the same data augmentation strategy with FixMatch for the OOD detection loss L_{OOD} to improve the generalization capability of the OOD-head. Using POT, we obtain the MSF for unlabeled samples based on the features of weakly-augmented versions of unlabeled samples. Further, we use the OOD loss to optimize the predicted OOD score for strongly-augmented versions of unlabeled samples. The overall computing framework is described in Algorithm 3.

[1]We use supervised and unsupervised losses from FixMatch to validate our approach. Additionally, in the ablation experiment, we apply our method to various other SSL techniques.

Algorithm 3 OSSL Computing Framework

Require: Feature extractor f, prediction head g, and OOD head h. Weight of unsupervised loss λ_u, weight of OOD loss λ_{OOD}. Weak data augmentation α and strong data augmentation $\mathcal{A}$. Entropy regularization parameter ε, and mass redundancy parameter k. Sample a batch of labeled data D_L and a batch of unlabeled data D_U
Feature Extractor:
Feature of the labeled batch $\mathbf{L} = f(D_L) \in \mathbb{R}^{n\times d}$
Feature of weakly augmented unlabeled batch $\mathbf{U} = f(\alpha(D_U)) \in \mathbb{R}^{m\times d}$
OOD Score:
OOD score of unlabeled batch using Algorithm2:
$Score_{\mathcal{U}} = \text{POD}(\mathbf{L}, \mathbf{U}, \varepsilon, k)$
Set the OOD score of labeled samples: $Score_{\mathcal{L}} = \mathbf{1}_n$
$Score = \text{CONCAT}(Score_{\mathcal{U}}, Score_{\mathcal{L}})$
Predict OOD Score:
The score of labeled batch: $Pred_{\mathcal{L}} = h(f(D_L))$
The score of strongly augmented unlabeled batch:
$Pred_{\mathcal{U}} = h(f(\mathcal{A}(D_U)))$
$Pred = \text{CONCAT}(Pred_{\mathcal{U}}, Pred_{\mathcal{L}})$
Calculate Loss:
calculate supervised loss of FixMatch L_x, unsupervised loss of FixMatch L_u, and L_{OOD} using (11)
loss = $L_x + \lambda_u L_u + \lambda_{\text{OOD}} L_{\text{OOD}}$

# of Labeled Classes	50		100		400	
Metric	Acc	AUROC	Acc	AUROC	Acc	AUROC
FixMatch	91.7±1.1	37.7±0.6	92.9±0.7	39.8±0.5	93.4±0.3	40.9±0.6
MTCF	79.7±0.9	96.6±0.5	86.3±0.9	98.2±0.3	91.0±0.5	98.9±0.1
T2T	88.2±0.7	75.5±0.5	89.0±1.0	77.2±0.2	90.3±0.5	82.3±0.2
OpenMatch	89.6±0.9	99.3±0.3	92.9±0.5	**99.7±0.2**	**94.1±0.5**	99.3±0.2
Ours	**92.1±0.2**	**99.7±0.1**	**92.9±0.2**	99.5±0.1	93.6±0.1	**99.4±0.1**

Table 1: Closed-set Accuracy Acc and Open-set Classification Quality AUROC on CIFAR10 dataset.

# of Known	55				80			
# of Labeled	50		100		50		100	
Metric	Acc	AUROC	Acc	AUROC	Acc	AUROC	Acc	AUROC
FixMatch	78.2±0.7	57.3±1.1	80.8±0.6	56.7±1.2	75.3±0.6	48.7±0.9	78.1±0.5	47.3±0.7
MTCF	66.5±1.2	81.2±3.4	72.1±0.5	80.7±4.6	59.9±0.8	79.4±2.5	66.4±0.3	73.2±3.5
T2T	72.2±1.4	60.4±1.6	73.1±0.8	59.8±1.4	63.5±1.2	55.0±1.8	66.8±0.7	55.4±1.5
OpenMatch	72.3±0.4	87.0±1.1	75.9±0.6	86.5±2.1	66.6±0.2	86.2±0.6	70.5±0.3	86.8±1.4
Ours	**78.7±0.2**	**88.4±0.1**	**81.1±0.1**	**89.5±0.3**	**75.4±0.1**	**88.1±0.3**	**78.1±0.1**	**88.0±0.1**

Table 2: Closed-set Accuracy Acc and Open-set Classification Quality AUROC on CIFAR100 dataset.

4 Experimental Results

Metrics. To evaluate the performance of OSSL methods, we employ closed-set classification accuracy to test the performance concerning the known classes. We use AUROC to assess the model's open-set classification ability in distinguishing between inliers and outliers.

Baselines. In terms of OSSL baselines, we evaluate our approach in comparison to various existing methods, including FixMatch, MTCF, T2T, OpenMatch, and IOMatch. Additionally, since our method is an extension of SSL, we also include FixMatch, a powerful and widely adopted method in SSL, as a baseline for reference. To assess the potential of FixMatch for OOD detection, we employ the maximum softmax prediction probability as the score function at test time.

4.1 Performance Evaluation

Results on CIFAR10 and CIFAR100. We evaluate the performance of POT in comparison to baselines on the widely used benchmark datasets for SSL, namely CIFAR10 and CIFAR100. In these experiments, we adapt a randomly initialized Wide ResNet-28-2 [Zagoruyko and Komodakis, 2016] with 1.5M parameters, in consistency with existing works. In particular, for CIFAR10, we divide it into 6 known classes and 4 unknown classes. For CIFAR100, we consider two settings: one with 80 known classes and 20 unknown classes, and another with 55 known classes and 45 unknown classes, organized according to superclasses. Note that we use the same set of hyper-parameters for all experiments[2].

Tables 1 and 2 report the results on CIFAR10 and CIFAR100, respectively. On the relatively simple CIFAR-10 dataset, our approach achieves state-of-the-art (SOTA) performance in terms of both closed-set accuracy ACC and open-set quality AUROC. Upon transitioning to the more challenging CIFAR-100 dataset, our method excels, dominating other OSSL competitors in both closed-set accuracy and AUROC. Notably, our approach showcases a noteworthy increase of more than 5% in closed-set accuracy. This substantiates the efficacy of our OOD binary detector, particularly in challenging and complex scenarios.

Metric	FixMatch	MTCF	T2T	OpenMatch	Ours
Acc	91.7±0.5	86.4±0.7	87.8±0.9	89.6±1.0	**92.0±0.3**
AUROC	45.1±1.2	93.8±0.8	55.7±10.8	96.4±0.7	**97.4±0.4**

Table 3: Performance Evaluation on ImageNet-30 dataset.

Results on ImageNet-30. We further evaluate our proposal in more complex and challenging settings. In particular, we use ImageNet-30 dataset [Hendrycks and Gimpel, 2016], a subset of ImageNet, comprising 30 classes. We select the first 20 classes in alphabetical order as known classes, while the remaining 10 classes are designated as unknown classes. We employ the ResNet-18 [He *et al.*, 2016] as the backbone network, the remaining experimental settings are the same as those conducted on the CIFAR dataset. Notably, as shown in Table 3, POT achieves the SOTA performance on the ImageNet-30 dataset, both in terms of inlier accuracy and AUROC.

Comparison with IOMatch. IOMatch treats all OOD samples as a distinct category, framing the OSSL task as an open-set classification challenge. To enable a direct comparison with IOMatch, we configure our method to output the model-predicted category for samples with an OOD score less than 0.5, and samples with a score exceeding 0.5 are directly classified as OOD. This alignment allows for a meaningful

[2] $\{\mu = 2, B = 64, N_e = 512, N_i = 1024, \varepsilon = 0.05, k = 2, \lambda_{\text{OOD}} = 0.01\}$. B indicates the batch size and μ is the relative size of batch size for unlabeled data. N_e indicates the total number of training epochs and N_i is the number of iterations per epoch.

Dataset	CIFAR10	CIFAR100		
Class Splits	6/4	20/80	50/50	80/20
IOMatch	93.9	67.3	69.8	64.8
Ours	**94.2**	**70.3**	**70.0**	**73.8**

Table 4: Closed-set Classification Accuracy on CIFAR.

Dataset	CIFAR10	CIFAR100		
Class Splits	6/4	20/80	50/50	80/20
IOMatch	79.0	58.5	60.8	54.4
Ours	**89.6**	**67.9**	**68.0**	**70.4**

Table 5: Open-set Classification Balanced Accuracy on CIFAR.

comparison using identical evaluation indicators.

Note that IOMatch addresses the OSSL problem with the constraints of limited labels, particularly in few-shot learning scenarios. In contrast, our method is designed for general semi-supervised learning scenarios. In the comparative study, we experimented with a setting featuring 25 labeled samples per class, in contrast to IOMatch's few-shot learning scenario with only 4 labeled samples per class. This intentional deviation in experimental design ensures a more comprehensive and representative assessment of our method's performance in semi-supervised learning contexts.

In Tables 4 and 5, it shows that our approach clearly demonstrates a substantial advantage in open-set classification. In closed-set classification, specifically with 80 known classes, our method outperforms IOMatch by over 9% in closed-set classification tasks. These findings underscore the significant strengths of our approach in effectively handling a wide spectrum of classification categories.

Effect of k and ε. Table 6 shows the performance under 9 different configurations of POT parameters. Notably, when the unlabeled dataset in the CIFAR10 dataset contains 40 percent of OOD samples, the model achieves excellent results with k ranging from 1.5 to 2.5. This observation underscores the validity of our method, particularly the strategy of generating redundant mass for unlabeled distributions.

Further, we examine the effect of the entropy regularization parameter on the model. We find that the model exhibits increased stability with larger values of ε. In addition, based on the results in Table 6, the model consistently outperforms most parameter combinations, demonstrating the robust generalization capabilities inherent in our proposed method.

k	1.5		2		2.5	
ε	Acc	AUROC	Acc	AUROC	Acc	AUROC
0.01	91.9	99.5	92.5	99.5	92.3	99.1
0.05	92.1	99.2	92.1	99.7	92.2	99.6
0.1	91.9	99.6	91.9	99.1	91.9	99.5

Table 6: Effect of k and ε.

Effect of λ_{OOD}. The impact of varying the weight assigned to the OOD loss is demonstrated in Table 7. The result shows that our approach yields superior performance when tuning the OOD loss weights to align with the order of magnitude of the supervised loss (e.g., weights of 0.1 or 0.01). This suggests that the OOD loss should have a balanced weight with the supervised loss, neither dominating it nor being dominated by it. In practice, since the supervised loss typically carries a smaller magnitude, it suggests to use small weights for the OOD loss when implementing our method.

# of Labeled	50		100		400	
λ_{OOD}	Acc	AUROC	Acc	AUROC	Acc	AUROC
1	91.1	98.6	92.6	96.5	93.6	99.0
0.1	91.9	99.0	92.7	99.1	**93.7**	99.3
0.01	**92.1**	**99.7**	**92.9**	**99.5**	93.6	**99.4**
0.001	91.7	98.0	92.6	98.5	93.9	99.0
0	91.7	37.7	92.9	39.8	93.4	40.9

Table 7: Effect of Different Weights of OOD loss.

Training Efficiency. When computing the MSF, POT demonstrates a per-iteration complexity of $O(mn)$, which is notably lower than the computational cost of conventional SSL methods. In our experiments on CIFAR-10 dataset, which includes 50 labeled samples per class and is executed on a single GPU, our method completes one training epoch in a total of 123 seconds, in which the computation of MSF takes only 6 seconds. It shows that our method achieves competitive performance in terms of efficiency.

Adaption to other SSL Methods. The OOD detector we introduced possesses the ability to seamlessly integrate with various off-the-shelf SSL methods, adapting them to OSSL scenarios. Table 8 illustrates the effectiveness of integrating our proposed OOD detector with FlexMatch [Zhang *et al.*, 2021], an advanced algorithm derived from FixMatch. These findings underscore the generality of our method to SSLs.

Dataset	CIFAR10		CIFAR100			
Class Split	6/4		55/45		80/20	
Metric	Acc	AUROC	Acc	AUROC	Acc	AUROC
FlexMatch	92.8	99.6	77.1	87.1	74.0	87.4
FixMatch	92.1	99.7	78.7	88.4	75.4	88.1

Table 8: Adaption to FlexMatch.

5 Conclusion

In this work, we introduce a novel approach for OOD detection within the context of OSSL. In particular, we formulate the OOD detection in OSSL as a POT problem. With POT theory, we propose a novel mass score function based on transport mass to measure the likelihood of a sample being an outlier. More, to train the model in an end-to-end manner, we devise an OOD loss, enabling the adaption of off-the-shelf SSL methods to OSSL settings via multi-task learning. Extensive experiments on multiple datasets and OSSL settings are conducted to evaluate our proposals.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61772492, Grant 62072428, Grant 62271465, in part by the Open Fund Project of Guangdong Academy of Medical Sciences, China under Grant YKY-KF202206, and in part by the Suzhou Basic Research Program under Grant SYG202338. Xike Xie is the corresponding author.

References

[Bauschke and Lewis, 2000] Heinz H Bauschke and Adrian S Lewis. Dykstras algorithm with bregman projections: A convergence proof. *Optimization*, 48(4):409–427, 2000.

[Benamou *et al.*, 2015] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[Bickel *et al.*, 2009] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(9), 2009.

[Caffarelli and McCann, 2010] Luis A Caffarelli and Robert J McCann. Free boundaries in optimal transport and monge-ampere obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[Caruana, 1997] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997.

[Chapel *et al.*, 2020] Laetitia Chapel, Mokhtar Z Alaya, and Gilles Gasso. Partial optimal tranport with applications on positive-unlabeled learning. *Advances in Neural Information Processing Systems*, 33:2903–2913, 2020.

[Chen *et al.*, 2020] Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3569–3576, 2020.

[Courty *et al.*, 2014] Nicolas Courty, Remi Flamary, Alain Rakotomamonjy, and Devis Tuia. Optimal transport for domain adaptation. In *NIPS 2014, Workshop on Optimal Transport and Machine Learning*, 2014.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[Dykstra, 1983] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[Fan *et al.*, 2023] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Ssb: Simple but strong baseline for boosting performance of open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16068–16078, 2023.

[Feng *et al.*, 2023] Chuanwen Feng, Yilong Ren, and Xike Xie. Ot-filter: An optimal transport filter for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16164–16174, 2023.

[Figalli and Glaudo, 2021] Alessio Figalli and Federico Glaudo. *An invitation to optimal transport, Wasserstein distances, and gradient flows*. 2021.

[Guo *et al.*, 2020] Lan-Zhe Guo, Zhen-Yu Zhang, Yuan Jiang, Yu-Feng Li, and Zhi-Hua Zhou. Safe deep semi-supervised learning for unseen-class unlabeled data. In *International Conference on Machine Learning*, pages 3897–3906. PMLR, 2020.

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

[He *et al.*, 2022] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Safe-student for safe deep semi-supervised learning with unseen-class unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14585–14594, 2022.

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

[Huang *et al.*, 2021] Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8310–8319, 2021.

[Laine and Aila, 2016] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[Lee and others, 2013] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

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

[Li *et al.*, 2023] Zekun Li, Lei Qi, Yinghuan Shi, and Yang Gao. Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15870–15879, 2023.

[Liu *et al.*, 2020] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[Lu *et al.*, 2023] Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, and Yang Cao. Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3291, 2023.

[Ma *et al.*, 2023] Qiankun Ma, Jiyao Gao, Bo Zhan, Yunpeng Guo, Jiliu Zhou, and Yan Wang. Rethinking safe semi-supervised learning: Transferring the open-set problem to a close-set one. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16370–16379, 2023.

[Saito *et al.*, 2021] Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. *arXiv preprint arXiv:2105.14148*, 2021.

[Sohn *et al.*, 2020] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[Sun *et al.*, 2022] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors, 2022.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[Villani, 2008] C Villani. Optimal transport, old and new. notes for the 2005 saint-flour summer school. *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer*, 3, 2008.

[Villani, 2021] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[Wang *et al.*, 2023] Yu Wang, Pengchong Qiao, Chang Liu, Guoli Song, Xiawu Zheng, and Jie Chen. Out-of-distributed semantic pruning for robust semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23849–23858, 2023.

[Xie *et al.*, 2020] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training, 2020.

[Xu *et al.*, 2020] Renjun Xu, Pelen Liu, Yin Zhang, Fang Cai, Jindong Wang, Shuoying Liang, Heting Ying, and Jianwei Yin. Joint partial optimal transport for open set domain adaptation. In *IJCAI*, pages 2540–2546, 2020.

[Yang *et al.*, 2023a] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9):8934–8954, September 2023.

[Yang *et al.*, 2023b] Yucheng Yang, Xiang Gu, and Jian Sun. Prototypical partial optimal transport for universal domain adaptation. 2023.

[Yu *et al.*, 2020] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[Zhang *et al.*, 2021] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021.