OTMatch: Improving Semi-Supervised Learning with Optimal Transport

Zhiquan Tan [1] **Kaipeng Zheng** [2] **Weiran Huang** [† 2 3]

Abstract

Semi-supervised learning has made remarkable strides by effectively utilizing a limited amount of labeled data while capitalizing on the abundant information present in unlabeled data. However, current algorithms often prioritize aligning image predictions with specific classes generated through self-training techniques, thereby neglecting the inherent relationships that exist within these classes. In this paper, we present a new approach called OTMatch, which leverages semantic relationships among classes by employing an optimal transport loss function to match distributions. We conduct experiments on many standard vision and language datasets. The empirical results show improvements in our method above baseline, this demonstrates the effectiveness and superiority of our approach in harnessing semantic relationships to enhance learning performance in a semi-supervised setting.

1. Introduction

Semi-supervised learning occupies a unique position at the intersection of supervised learning and self-supervised learning paradigms (Tian et al., 2020; Chen et al., 2020a). The fundamental principle behind semi-supervised learning lies in its ability to leverage the latent patterns and structures of a substantial amount of unlabeled samples to collaborate with conventional supervised learning on labeled samples. It has demonstrated remarkable performance without the need for extensive human efforts in data collection (Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2022d).

Pseudo-labeling-based methods have dominated the research in semi-supervised learning. It dynamically assigns labels to unlabeled samples to prepare an extended dataset with labeled samples for model training (Lee et al., 2013; Tschannen et al., 2019; Berthelot et al., 2019b; Xie et al., 2020; Sohn et al., 2020; Gong et al., 2021; Zhang et al., 2021; Wang et al., 2022d), allowing the model to benefit from the potential knowledge contained in these unlabeled samples. Typically, pseudo-labels are derived from the neural network's confidence, and the cross-entropy loss is used to align the prediction on the strongly augmented image view with the generated pseudo-labels. Following this paradigm, recent works have achieved state-of-the-art performance (Zhang et al., 2021; Wang et al., 2022d) in semi-supervised learning. However, it has been demonstrated that the overconfidence of models can be observed when assigning pseudo labels during training (Chen et al., 2022). This means that the model could confidently assign samples to incorrect categories, due to the limited labeled samples in semi-supervised learning. Such unreliability in confidence is particularly pronounced when dealing with extremely limited labeled samples. As a result, those misclassified samples can misguide the model's learning process, resulting in a decrease in performance.

In this work, we propose a novel solution to alleviate the issue of overconfidence in the model in pseudo-labeling-based methods by incorporating inter-class semantic relationships. Particularly, we find that a key issue lies in simply employing the traditional cross-entropy loss to train the model with the pseudo-labeled samples. In the cross-entropy loss, the representation of a sample is aligned with a single category. However, when a sample is assigned an incorrect pseudo-label, this leads the model to learn the representation in the wrong direction. To address this issue, for pseudo-labeled samples, we propose to incorporate comprehensive inter-class relationships for regularization, instead of a single-category target, to guide model training, hence offering improved robustness. Particularly, we use optimal transport to tackle this problem as the predicted probability and pseudo labels are naturally two distributions that need to be matched, and semantic information can be injected into the cost matrix in optimal transport loss.

The paper is organized in the following manner: We present a new perspective on current semi-supervised learning methods that rely on pseudo-labeling. We view these methods as aiming to align the semantic distribution captured by the teacher and student models, motivated by (Shi et al.,

[1]Department of Mathematical Sciences, Tsinghua University [2]MIFA Lab, Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University [3]Shanghai AI Laboratory. [†]Correspondence to: Weiran Huang <weiran.huang@outlook.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

2023). We further improve this by proposing a novel training algorithm OTMatch, where we bootstrap the cost in optimal transport from the class embedding throughout model training, thus fully capitalizing on semantic relationships between classes. The diagram of OTMatch is presented in Figure 1.

Our contributions can be summarized as follows:

- We provide a novel understanding of current pseudo-labeling-based semi-supervised learning methods by viewing them as matching the distribution of semantics obtained by the teacher and student models using (inverse) optimal transport. We also extend this framework to analyze algorithms in self-supervised learning.
- We propose OTMatch, a novel semi-supervised learning algorithm that exploits the semantic relationship between classes to alleviate the issue of model overconfidence caused by limited labeled samples.
- We carry out experiments on well-known vision datasets such as CIFAR 10/100, STL-10, and ImageNet, noting that our method shows improvements, particularly in challenging situations with very few labeled samples. Additionally, we perform experiments in the language modality, discovering that our method is effective there as well.

2. Related Work

Semi-supervised learning, aiming to enhance model performance through the utilization of abundant unlabeled data, has attracted considerable attention in recent years (Chen et al., 2020b; Assran et al., 2021; Wang et al., 2021; Zhang et al., 2023b; Chen et al., 2023b; Nassar et al., 2023; Huang et al., 2023). The invariance principle serves as the foundation for many effective semi-supervised algorithms. Essentially, this principle posits that two semantically similar images should yield similar representations when processed by a same backbone.

Consistency regularization. Consistency regularization, initially introduced in the Π-Model (Rasmus et al., 2015b), has emerged as a prevalent technique for implementing the invariance principle. This method has gained widespread adoption in subsequent research (Tarvainen & Valpola, 2017; Laine & Aila, 2016; Berthelot et al., 2019b). Consistency regularization entails the generation of pseudo-labels and the application of appropriate data augmentation strategies (Tschannen et al., 2019; Berthelot et al., 2019b; Xie et al., 2020; Sohn et al., 2020; Gong et al., 2021). Pseudo-labels are created for unlabeled data and utilized in subsequent training iterations (Lee et al., 2013). The conventional approach involves minimizing the cross-entropy objective to align the predicted pseudo-labels of two distorted images, typically obtained through data augmentation (Rasmus et al., 2015b; Laine & Aila, 2016; Tarvainen & Valpola, 2017). Extensive research has recently focused on generating efficient and informative pseudo-labels (Hu et al., 2021; Nassar et al., 2021; Xu et al., 2021; Zhang et al., 2021; Li et al., 2022; Wang et al., 2022b), achieving state-of-the-art performance. SimMatch (Zheng et al., 2022) and CoMatch (Li et al., 2021) also investigate contrastive learning for consistency regularization. The efficacy of consistency regularization has been demonstrated as a simple yet effective approach, serving as a foundational component in numerous state-of-the-art semi-supervised learning algorithms.

Improving pseudo-label quality. In the realm of semi-supervised learning, the current discourse surrounding consistency regularization primarily revolves around augmenting the quality of pseudo-labels. SimPLE (Hu et al., 2021) introduces a paired loss function that diminishes the statistical discrepancy between confident and analogous pseudo-labels, thereby enhancing their quality. Dash (Xu et al., 2021) and FlexMatch (Zhang et al., 2021) propose dynamic and adaptable filtering techniques for pseudo-labels, which are better suited for the training process. CoMatch (Li et al., 2021) advocates for the integration of contrastive learning within the framework of semi-supervised learning, enabling the simultaneous learning of two representations of the training data. SemCo (Nassar et al., 2021) takes into account external label semantics to safeguard against pseudo-label quality deterioration for visually similar classes, employing a co-training approach. FreeMatch (Wang et al., 2022d) proposes a self-adjusting confidence threshold that considers the learning status of the models, allowing for improved control over pseudo-label quality. MaxMatch (Li et al., 2022) presents a consistency regularization technique that minimizes the most substantial inconsistency between an original unlabeled sample and its multiple augmented versions, accompanied by theoretical guarantees. NP-Match (Wang et al., 2022a) employs neural processes to amplify the quality of pseudo-labels. SEAL (Tan et al., 2023a) introduces a methodology that facilitates the concurrent learning of a data-driven label hierarchy and the execution of semi-supervised learning. SoftMatch (Chen et al., 2023a) addresses the inherent trade-off between the quantity and quality of pseudo-labels by utilizing a truncated Gaussian function to assign weights to samples based on their confidence.

Unlike previous works focusing on enhancing pseudo-label quality, we address the issue of overconfidence in models from an orthogonal perspective by incorporating inter-class relationships as constraints. Taherkhani et al. (2020); Tai et al. (2021) also explore using optimal transport in semi-supervised learning. However, they still use optimal transport for pseudo-label filtering. In contrast, we employ optimal transport theory to provide a novel understanding of current pseudo-labeling methods. We further improve

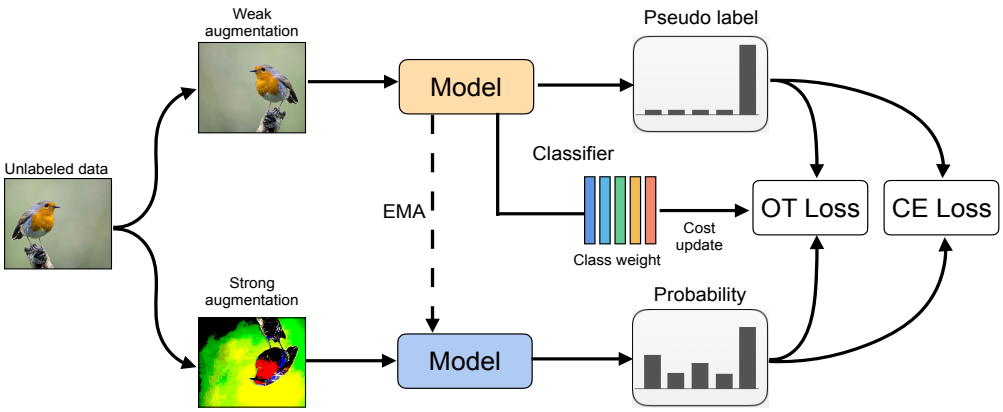

Figure 1. To obtain a pseudo-label, a model is fed with a weakly augmented image. Then, the model predicts the probability of a strongly augmented version of the same image. The loss includes cross-entropy and optimal transport loss, which considers the probability and pseudo-label. The cost used in optimal transport is adjusted based on the model's classification head weight.

this by extracting inter-class semantic relationships from the model's learning dynamic to update the cost matrix, which has never been explored before.

3. Preliminary

3.1. Problem setting and notations

Throughout a semi-supervised learning process, it is customary to have access to both labeled and unlabeled data. Each batch is a mixture of labeled data and unlabeled data. Assume there are B labeled samples $\{(\mathbf{x}_{l_i}, \mathbf{y}_{l_i})\}_{i=1}^{B}$ and μB unlabeled samples $\{\mathbf{x}_{u_i}\}_{i=1}^{\mu B}$ in a mixed batch, where μ is the ratio of unlabeled samples to labeled samples. We adopt the convention in semi-supervised learning (Zhang et al., 2021; Wang et al., 2022d) that there will be a teacher and student network that shares the same architecture. The teacher network does not update through gradient but by exponential moving average (EMA) instead. There will also be two sets of augmentations of different strengths, namely weak augmentation $\omega(\cdot)$ and strong augmentation $\Omega(\cdot)$.

For labeled data, the loss is the classical cross-entropy loss as follows:

$$\mathcal{L}_{\text{sup}} = \frac{1}{B}\sum_{i=1}^{B} \text{H}(\mathbf{y}_{l_i}, \text{Pr}(\omega(\mathbf{x}_{l_i}))), \tag{1}$$

where $\text{Pr}(\omega(\mathbf{x}_{l_i}))$ denotes the output probability and $\text{H}(\cdot,\cdot)$ denotes the cross-entropy loss.

The unsupervised loss L_{un} is usually the main focus of improving semi-supervised learning. FixMatch (Sohn et al., 2020) introduces the idea of using a fixed threshold τ to only assign pseudo labels to those samples with enough confidence. Later, a line of works like FlexMatch (Zhang et al., 2021) and FreeMatch (Wang et al., 2022d) seeks to improve the threshold selection strategy. This loss can be formally described as follows:

$$\mathcal{L}_{\text{un1}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbf{I}(\max(\mathbf{q}_{u_i}) > \tau)\text{H}(\hat{\mathbf{q}}_{u_i}, \mathbf{Q}_{u_i}), \tag{2}$$

where $\mathbf{q}_{u_i}$ is the probability of (the teacher) model on the weakly-augmented image, $\mathbf{Q}_{u_i}$ is the probability of (the student) model on the strongly-augmented image, and $\hat{\mathbf{q}}_{u_i}$ denotes the generated one-hot hard pseudo label.

FreeMatch (Wang et al., 2022d) also introduces a fairness loss to make the class distribution more balanced. The loss is given as follows:

$$\mathcal{L}_{\text{un2}} = -\text{H}(\text{SumNorm}(\frac{\mathbf{p}_1}{\mathbf{h}_1}), \text{SumNorm}(\frac{\mathbf{p}_2}{\mathbf{h}_2})), \tag{3}$$

where $\mathbf{p}_1$ and $\mathbf{h}_1$ denote the mean of model predictions and histogram distribution on weakly-augmented images respectively, $\mathbf{p}_2$ and $\mathbf{h}_2$ is defined on the pseudo-labeled strongly-augmented images. As the prediction on the weakly-augmented image is more accurate, using cross-entropy loss here mimics the maximization of entropy.

3.2. Optimal transport

The Kantorovich formulation of discrete optimal transport (Kantorovich, 1942), also known as the transportation problem, provides a mathematical definition for finding the optimal transportation plan between two discrete probability distributions. Let's consider two discrete probability distributions, denoted as μ and ν, defined on two finite sets of points, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m\}$ and $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, respectively. The goal is to find a transportation plan that

minimizes the total transportation cost while satisfying certain constraints. The transportation plan specifies how much mass is transported from each point in $\mathbf{X}$ to each point in $\mathbf{Y}$. This is achieved by defining a transportation (plan) matrix $\mathbf{T} = [\mathbf{T}_{ij}]$, where $\mathbf{T}_{ij}$ represents the amount of mass transported from point $\mathbf{x}_i$ to point $\mathbf{y}_j$.

For ease of notation, we present the definition in the following form:

$$\begin{aligned} &\min \quad \langle \mathbf{C}, \mathbf{T} \rangle \\ &\text{s.t. } \mathbf{T} \in U(\mu, \nu) = \{\mathbf{T} \in \mathbb{R}_+^{m \times n} \mid \mathbf{T}\mathbf{1}_n = \mu, \mathbf{T}^T \mathbf{1}_m = \nu\}, \end{aligned}$$

where $\mathbf{C}$ denotes the cost matrix and $\langle \mathbf{C}, \mathbf{T} \rangle = \sum_{ij} \mathbf{C}_{ij} \mathbf{T}_{ij}$ is the inner product between matrices.

In this formulation, $\mathbf{c}_{ij}$ represents the cost between point $\mathbf{x}_i$ and point $\mathbf{y}_j$. It could be any non-negative cost function that captures the transportation cost between the points. The objective is to minimize the total cost, which is the sum of the products of the transportation amounts $\mathbf{T}_{ij}$ and their corresponding costs $\mathbf{c}_{ij}$. We denote the optimal transport distance as $\mathcal{W}(\mu, \nu)$.

The constraints ensure that the transportation plan satisfies the conservation of mass: the total mass transported from each point in $\mathbf{X}$ should be equal to its mass in distribution μ, and the total mass received at each point in $\mathbf{Y}$ should be equal to its mass in distribution ν. Additionally, the transportation amounts $\mathbf{T}_{ij}$ are non-negative.

The solution to this optimization problem provides the optimal transportation plan, which specifies how much mass is transported from each point in $\mathbf{X}$ to each point in $\mathbf{Y}$ to minimize the total cost. Algorithms, such as the Hungarian algorithm (Kuhn, 1955) can be applied to solve this problem with the complexity of $O(m^2 n)$.

The computational complexity to solve the general optimal transport problem is relatively high. Cuturi (2013) proposes an entropic regularized version of the optimal transport problem as follows:

$$\begin{aligned} \min \quad & \langle \mathbf{C}, \mathbf{T} \rangle - \epsilon \mathrm{H}(\mathbf{T}) \\ \text{subject to} \quad & \mathbf{T} \in U(\mu, \nu), \end{aligned}$$

where $\epsilon > 0$ is a hyperparameter and $\mathrm{H}(\mathbf{T}) = \sum_{i,j}(1 - \log \mathbf{T}_{ij})\mathbf{T}_{ij}$.

This regularized problem can be solved by the Sinkhorn algorithm efficiently with a complexity of $O(\frac{mn}{\epsilon})$. It can be shown that this regularized version approximately solves the initial discrete optimal transport problem.

4. Understanding FreeMatch via Optimal Transport

In this section, we start by using the view of semantic matching of distributions to understand the state-of-the-art pseudo-labeling-based semi-supervised learning methods. The main tool we will use is the optimal transport. Without loss of generality, we take the state-of-the-art method FreeMatch (Wang et al., 2022d) as an example. For simplicity, we abbreviate the exponential moving average (EMA) operation.

We find that a natural semantic distribution naturally arises in (semi-) supervised learning, which is the semantic distribution between samples and classes. Specifically, denote $\mu = \frac{1}{m}\mathbf{1}_m$, then the following set of matrices can be seen as the semantic distribution between m samples and K classes (Because each row of $\mathbf{T}$ captures the semantic of the sample's relationship to each class, the bigger the semantic distribution, the bigger the values.):

$$U(\mu) = \{\mathbf{T} \in \mathbb{R}_+^{m \times K} \mid \mathbf{T}\mathbf{1}_K = \mu\},$$

where $\mathbf{1}_K$ is an all one vector.

The above observation is also noticed in the setting of supervised learning setting by (Shi et al., 2023). In the following, we will first recap the derivation in their paper for supervised loss, where (Shi et al., 2023) introduces the framework of inverse optimal transport (IOT).

IOT aims to infer the cost function from the observed empirical semantic distribution matrix. It usually parameterizes the cost matrix into a learnable matrix $\mathbf{C}^\theta$ and solves the following optimization problem:

$$\begin{aligned} \min \quad & \mathrm{KL}(\bar{\mathbf{T}} \| \mathbf{T}^\theta) \\ \text{subject to} \quad & \mathbf{T}^\theta = \underset{\mathbf{T} \in U(\mu)}{\arg\min} \langle \mathbf{C}^\theta, \mathbf{T} \rangle - \epsilon \mathrm{H}(\mathbf{T}), \end{aligned} \tag{4}$$

where $\bar{\mathbf{T}}$ is a given semantic distribution matrix and the KL divergence is defined as follows.

Definition 4.1. For any two positive measures (distributions) $\mathbf{P}$ and $\mathbf{Q}$ on the same support $\mathcal{X}$, the KL divergence from $\mathbf{Q}$ to $\mathbf{P}$ is given by:

$$\mathrm{KL}(\mathbf{P} \| \mathbf{Q}) = -\sum_{x \in \mathcal{X}} \mathbf{P}(x) \log \frac{\mathbf{P}(x)}{\mathbf{Q}(x)} - \sum_{x \in \mathcal{X}} \mathbf{P}(x) + \sum_{x \in \mathcal{X}} \mathbf{Q}(x). \tag{5}$$

Shi et al. (2023) show that

$$\mathbf{T}_{ij} = \frac{1}{m} \frac{\exp\left(-\mathbf{C}_{ij}/\epsilon\right)}{\sum_{k=1}^{K} \exp\left(-\mathbf{C}_{ik}/\epsilon\right)} \tag{6}$$

is the closed-form solution to the optimization problem (7):

$$\underset{\mathbf{T} \in U(\mu)}{\arg\min} \quad \langle \mathbf{C}, \mathbf{T} \rangle - \epsilon \mathrm{H}(\mathbf{T}). \tag{7}$$

Then consider a batch of labeled data is $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^B$, where $\mathbf{x}_i$ represents the i-th image in the dataset and $\mathbf{y}_i$ is the label for this image. We can construct a "ground truth" semantic distribution matrix $\bar{\mathbf{T}}$ by setting $\bar{\mathbf{T}}_{ij} = \frac{1}{B}\delta_j^{\mathbf{y}_i}$.

Denote the logits generated by the neural network for each image $\mathbf{x}_i$ as $\mathbf{l}_\theta(x_i)$. By setting the cost matrix $\mathbf{C}_{ij}^\theta = c - \mathbf{l}_\theta(x_i)_j$ (c is a large constant), simplifying the transport matrix (6) by dividing the same constant $\exp(-c/\epsilon)$, assuming there are a total of K labels, the transportation matrix is given by:

$$\mathbf{T}_{ij}^\theta = \frac{1}{m}\frac{\exp(\mathbf{l}_\theta(\mathbf{x}_i)_j/\epsilon)}{\sum_{k=1}^K \exp(\mathbf{l}_\theta(\mathbf{x}_i)_k/\epsilon)}. \tag{8}$$

It is then straightforward to find that the loss in problem (4) is reduced as follows:

$$\mathcal{L} = -\sum_{i=1}^B \log \frac{\exp(\mathbf{l}_\theta(x_i)_j/\epsilon)}{\sum_{k=1}^K \exp(\mathbf{l}_\theta(\mathbf{x}_i)_k/\epsilon)} + \text{Const},$$

which exactly mirrors the supervised cross-entropy loss in semi-supervised learning with a temperature parameter ϵ.

Next, we delve into comprehending the more challenging unsupervised loss. We introduce a lemma that is very useful afterward in analyzing the unsupervised loss.

Lemma 4.2. *$\frac{\sum_{i=1}^m s_i}{m}$ is the unique solution of the optimization problem:*

$$\min_x \mathcal{W}(\delta_x, \sum_{i=1}^m \frac{1}{m}\delta_{s_i}),$$

where the underlying cost is the square of l^2 distance.

FreeMatch employs an adaptive threshold for pseudo-labeling. It is essentially equivalent to generating the threshold based on the semantic distribution matrix of (the teacher) model from equation (8). In particular, we start by analyzing the global threshold in FreeMatch, which aims to modulate the global confidence across different classes. Given that, each row of the matching matrix indicates the estimated probability $\mathbf{q}_{u_i}$ over the K classes. An intuitive idea is to associate each unlabeled sample u_i with a real number indicating the prediction confidence, thus we can take $\max(\mathbf{q}_{u_i})$ as a representative. Consequently, the general prediction confidence over the unlabeled data can be represented by a probability distribution $\sum_{i=1}^{\mu B}\frac{1}{\mu B}\delta_{\max(\mathbf{q}_{u_i})}$ that captures the full knowledge of the predictions. By identifying the global threshold τ with a probability distribution δ_τ and using Lemma 4.2, the global threshold can be calculated as

$$\tau = \frac{\sum_{i=1}^{\mu B}\max(\mathbf{q}_{u_i})}{\mu B}.$$

Since the global threshold does not accurately reflect the learning status of each class, we can refine the global threshold by incorporating the learning information for each class. Note the prediction $\mathbf{q}_{u_i}$ not only provide the "best" confidence $\max(\mathbf{q}_{u_i})$, but also suggest the confidence on each class k ($1 \le k \le K$).

By aggregating all confidences of unlabeled data for class k and organizing them into a probability distribution $\sum_{i=1}^{\mu B}\frac{1}{\mu B}\delta_{\mathbf{q}_{u_i}(k)}$, we can calculate the local importance $\mathbf{p}_1(k) = \sum_{i=1}^{\mu B}\frac{\mathbf{q}_{u_i}(k)}{\mu B}$ using a similar argument in the global threshold case. By adjusting the relative threshold according to the importance, we can finally derive the (local) threshold as follows:

$$\tau(k) = \frac{\mathbf{p}_1(k)}{\max_{k'}\mathbf{p}_1(k')}\frac{\sum_{i=1}^{\mu B}\max(\mathbf{q}_{u_i})}{\mu B}.$$

When dealing with unlabeled data, there is no "ground truth" semantic distribution matrix like the supervised cases. Therefore, we use the semantic distribution matrix **after** threshold filtering to serve as a "ground truth". It's important to highlight that both the teacher and student models share a similar format of the semantic distribution matrix given by equation (8). Consequently, unlabeled samples filtered by the teacher model are also excluded from the student model to avoid transferring predictions with low confidence. We use different ϵ for the teacher and student models (student use $\epsilon = 1$) as they have different confidences, when the ϵ for the teacher model is approaching zero, equation (8) will recover the one-hot pseudo label.

Thus the reduced semantic distribution matrices of teacher and student models are no longer probability matrices. As they still form positive matrices, by converting each row of the teacher model's predictions into pseudo labels and using the definition of KL divergence in equation (5) for positive distributions (measures). We find that the KL divergence between the teacher and student semantic distribution matrices recovers exactly the unsupervised loss $\mathcal{L}_{\text{un1}}$. The fairness loss $\mathcal{L}_{\text{un2}}$ can be similarly understood according to Lemma 4.2.

Remark: More algorithms derived from the framework of using optimal transport to match semantics can be found in Appendix B.

5. OTMatch: Improving Semi-Supervised Learning with Optimal Transport

From the derivation in section 4, we found that the performance of student models depends on the semantic distribution given by the teacher model. As teacher model generates one-hot pseudo labels, which makes it more likely to be overconfident in its predictions. This overconfidence leads to misclassifications and hampers the model's performance.

Algorithm 1 OTMatch training algorithm at t-th step

Input: Number of classes K, labeled samples $\{(\mathbf{x}_{l_i}, \mathbf{y}_{l_i})\}_{i=1}^{B}$, unlabeled samples $\{\mathbf{x}_{u_i}\}_{i=1}^{\mu B}$, FreeMatch loss weights w_1, w_2, and EMA decay m, OT loss balancing weight λ, normalized classification head vectors $\{\mathbf{v}_i\}_{i=1}^{K}$.
FreeMatch loss:
Calculate $\mathcal{L}_{\text{sup}}$ using equation (1)
$\tau_t = m\tau_{t-1} + (1-m)\frac{1}{\mu B}\sum_{i=1}^{\mu B} \max(\mathbf{q}_{u_i})$
$\tilde{p}_t = m\tilde{p}_{t-1} + (1-m)\frac{1}{\mu B}\sum_{b=1}^{\mu B} \mathbf{q}_{u_i}$
$\tilde{h}_t = m\tilde{h}_{t-1} + (1-m)\,\text{Hist}_{\mu B}\left(\hat{\mathbf{q}}_{u_i}\right)$
for $c = 1$ to K **do**
  $\tau_t(c) = \text{MaxNorm}(\tilde{p}_t(c)) \cdot \tau_t$
end for
Calculate $\mathcal{L}_{\text{un1}}$ using equation (2)
$\overline{p} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbf{I}\left(\max\left(\mathbf{q}_{u_i}\right) \geq \tau_t(\arg\max\left(\mathbf{q}_{u_i}\right)\right)\mathbf{Q}_{u_i}$
$\overline{h} = \text{Hist}_{\mu B}(\mathbf{I}\left(\max\left(\mathbf{q}_{u_i}\right) \geq \tau_t(\arg\max\left(\mathbf{q}_{u_i}\right)\right)\hat{\mathbf{Q}}_{u_i}))$
Calculate $\mathcal{L}_{\text{un2}}$ using equation (3)
$\mathcal{L}_{\text{FreeMatch}} = \mathcal{L}_{\text{sup}} + w_1\mathcal{L}_{\text{un1}} + w_2\mathcal{L}_{\text{un2}}$
Cost update:
$\mathbf{C}_t(i,j) = m\mathbf{C}_{t-1}(i,j) + (1-m)(1 - \langle \mathbf{v}_i, \mathbf{v}_j\rangle)$
OT loss:
$\mathcal{L}_{\text{un3}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbf{I}(\max(\mathbf{q}_{u_i}) > \tau_t(\arg\max(\mathbf{q}_{u_i})))$
$\sum_{k=1}^{K} \mathbf{C}_t(\arg\max(\mathbf{q}_{u_i}), k)\mathbf{Q}_{u_i}(k)$
OTMatch loss:
$\mathcal{L}_{\text{OTMatch}} = \mathcal{L}_{\text{FreeMatch}} + \lambda\mathcal{L}_{un3}$

To tackle this issue, we propose a novel solution that incorporates inter-class semantic relationships to alleviate model overconfidence in pseudo-labeling-based methods. Optimal transport is also a suitable tool here because the pseudo-label and the predicted probability given by the student model are two distributions. By considering comprehensive inter-class relationships instead of relying on a single-category target, we aim to improve the model's robustness and accuracy.

The cost function in optimal transport plays an important role. Frogner et al. (2015) construct cost using additional knowledge like word embedding. This approach may not be problem-specific and incorporates additional knowledge. Unlike previous works, we propose that the cost matrix can actually be effectively bootstrapped from the model itself. The basic idea is to "infer" the cost from the learning dynamic of the model. Since the model parameters are updated from (the stochastic) gradient descent method, our initial step involves analyzing the gradients. To simplify the analysis, assume the feature extracted by the model as unconstrained variables. Suppose the last layer of the neural network weights are denoted by $\mathbf{W} = [\mathbf{w}_1\mathbf{w}_2\cdots\mathbf{w}_K]$. Define the predicted probability matching for the image $\mathbf{x}$ as follows:

$$\mathbf{p}_k(f_\theta(\mathbf{x})) = \frac{\exp\left(f_\theta(\mathbf{x})^T\mathbf{w}_k\right)}{\sum_{k'=1}^{K}\exp\left(f_\theta(\mathbf{x})^T\mathbf{w}_{k'}\right)}, 1 \leq k \leq K.$$

Then we can calculate the loss's gradient for each image embedding with a label (or pseudo label) k as follows:

$$\frac{\partial\mathcal{L}}{\partial f_\theta(\mathbf{x})} = -\left(1 - \mathbf{p}_k(f_\theta(\mathbf{x}))\right)\mathbf{w}_k + \sum_{k'\neq k}^{K}\mathbf{p}_{k'}(f_\theta(\mathbf{x}))\mathbf{w}_{k'},$$

where $\mathcal{L}$ is the supervised loss $\mathcal{L}_{sup}$ or unsupervised loss $\mathcal{L}_{un1}$.

As the goal is to push $f_\theta(\mathbf{x})$ to the direction of $\mathbf{w}_k$, the updated score $U(\mathbf{x})$ along $\mathbf{w}_k$ during SGD on $f_\theta(\mathbf{x})$ can be calculated as:

$$\begin{aligned} U(\mathbf{x}) &= \langle -\frac{\partial\mathcal{L}}{\partial f_\theta(\mathbf{x})}, \mathbf{w}_k\rangle \\ &= \left(1 - \mathbf{p}_k(f_\theta(\mathbf{x}))\right)\langle\mathbf{w}_k, \mathbf{w}_k\rangle - \sum_{k'\neq k}^{K}\mathbf{p}_{k'}(f_\theta(\mathbf{x}))\langle\mathbf{w}_{k'}, \mathbf{w}_k\rangle. \end{aligned} \tag{9}$$

Note $U(\mathbf{x})$ reflects the hardness of classifying image $\mathbf{x}$ into class k. Thus we would like our expected cost of classification $C(\mathbf{x})$ to be proportional to $U(\mathbf{x})$. Using the law of probability, we can decompose $C(\mathbf{x})$ as follows:

$$C(\mathbf{x}) = \mathbb{E}_k(\text{Cost} \mid \mathbf{x}) = \sum_{k'=1}^{K}\mathbf{C}_{kk'}p_{k'}(f_\theta(\mathbf{x})).$$

When $\|\mathbf{w}_i\|_2 = 1$, by setting $\mathbf{C}_{kk'} = 1 - \langle\mathbf{w}_k, \mathbf{w}_{k'}\rangle$ we can demonstrate that $C(\mathbf{x}) = U(\mathbf{x})$.

Hence, it is evident that inter-class semantic relationships are indeed the core of the construction of an effective cost. Taking the fluctuation of batch training into consideration, we finally derive the cost update formula as follows:

$$\mathbf{C}_{kk'} = m\mathbf{C}_{kk'} + (1-m)(1 - \langle\mathbf{v}_k, \mathbf{v}_{k'}\rangle),$$

where the cost is initialized by discrete metric, m is the momentum coefficient, and $\mathbf{v}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|_2}$.

The computation cost of calculating the optimal transport cost is relatively high, which hinders its application. We present a lemma that shows under some mild conditions in semi-supervised learning, optimal transport can be calculated in complexity $O(K)$.

Lemma 5.1. *Suppose two probability distributions μ and ν support on $\mathcal{X}$ and suppose $\mid\mathcal{X}\mid = K$. Suppose the cost is generated by a metric and there exists k such that $\mu(i) \leq \nu(i)$ for any $i \neq k$. Then $\mathcal{W}(\mu,\nu) = \sum_{i=1}^{K}\mathbf{C}_{ik}(\nu(i) - \mu(i))$.*

Table 1. Error rates (100% - accuracy) on CIFAR-10/100, and STL-10 datasets for state-of-the-art methods in semi-supervised learning. Bold indicates the best performance, and underline indicates the second best.

Dataset	CIFAR-10			CIFAR-100		STL-10	
# Label	10	40	250	400	2500	40	1000
Π Model (Rasmus et al., 2015a)	79.18±1.11	74.34±1.76	46.24±1.29	86.96±0.80	58.80±0.66	74.31±0.85	32.78±0.40
Pseudo Label (Lee et al., 2013)	80.21± 0.55	74.61±0.26	46.49±2.20	87.45±0.85	57.74±0.28	74.68±0.99	32.64±0.71
VAT (Miyato et al., 2018)	79.81± 1.17	74.66±2.12	41.03±1.79	85.20±1.40	46.84±0.79	74.74±0.38	37.95±1.12
MeanTeacher (Tarvainen & Valpola, 2017)	76.37± 0.44	70.09±1.60	37.46±3.30	81.11±1.44	45.17±1.06	71.72±1.45	33.90±1.37
MixMatch (Berthelot et al., 2019b)	65.76± 7.06	36.19±6.48	13.63±0.59	67.59±0.66	39.76±0.48	54.93±0.96	21.70±0.68
ReMixMatch (Berthelot et al., 2019a)	20.77± 7.48	9.88±1.03	6.30±0.05	42.75±1.05	**26.03±0.35**	32.12±6.24	6.74±0.17
UDA (Xie et al., 2020)	34.53± 10.69	10.62±3.75	5.16±0.06	46.39±1.59	27.73±0.21	37.42±8.44	6.64±0.17
FixMatch (Sohn et al., 2020)	24.79± 7.65	7.47±0.28	5.07±0.05	46.42±0.82	28.03±0.16	35.97±4.14	6.25±0.33
Dash (Xu et al., 2021)	27.28± 14.09	8.93±3.11	5.16±0.23	44.82±0.96	27.15±0.22	34.52±4.30	6.39±0.56
MPL (Pham et al., 2021)	23.55± 6.01	6.93±0.17	5.76±0.24	46.26±1.84	27.71±0.19	35.76±4.83	6.66±0.00
FlexMatch (Zhang et al., 2021)	13.85± 12.04	4.97±0.06	4.98±0.09	39.94±1.62	26.49±0.20	29.15±4.16	5.77±0.18
FreeMatch (Wang et al., 2022d)	8.07± 4.24	4.90±0.04	4.88±0.18	37.98±0.42	26.47±0.20	15.56±0.55	5.63±0.15
OTMatch (Ours)	**4.89± 0.76**	**4.72± 0.08**	**4.60± 0.15**	**37.29± 0.76**	26.04± 0.21	**12.10± 0.72**	**5.60± 0.14**

Thus we can finally obtain our optimal transport-based unsupervised loss as follows:

$$\mathcal{L}_{\text{un3}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B}\mathbf{I}(\max(\mathbf{q}_{u_i}) > \tau(\arg\max(\mathbf{q}_{u_i})))\mathbf{W}(\hat{\mathbf{q}}_{u_i}, \mathbf{Q}_{u_i}).$$

When combining our method with FreeMatch, we obtain our final OTMatch loss as $\mathcal{L}_{\text{OTMatch}} = \mathcal{L}_{\text{FreeMatch}} + \lambda\mathcal{L}_{\text{un3}}$, where λ is a hyperparameter. This loss considers incorporating semantic information to distinguish two distribution $\hat{\mathbf{q}}_{u_i}$ and $\mathbf{Q}_{u_i}$. The whole process of our method is outlined in Algorithm 1.

Interestingly, the loss $\mathcal{L}_{un3}$ can also be interpreted using the view of self-attention (Vaswani et al., 2017). Setting $f_\theta(\mathbf{x})$ as query, $\mathbf{w}_j$ $(1 \le j \le K)$ as keys and $\mathbf{v}_j$ $(1 \le j \le K)$ as values, recall the definition of self-attention, $\sum_{i=1}^{K}\mathbf{p}_i(f_\theta(\mathbf{x}))\mathbf{v}_i$ is exactly the representation generated by self-attention.

For an unlabeled image $\mathbf{x}$ with pseudo label k, the loss can be reformulated as: $\mathcal{W}(\delta_k, \Pr(\mathbf{x})) = \sum_{i=1}^{K}\mathbf{C}_{ik}\Pr(i \mid \mathbf{x}) = \sum_{i=1}^{K}(1 - \langle\mathbf{v}_i, \mathbf{v}_k\rangle)\mathbf{p}_i(f_\theta(\mathbf{x})) = 1 - \langle\sum_{i=1}^{K}\mathbf{p}_i(f_\theta(\mathbf{x}))\mathbf{v}_i, \mathbf{v}_k\rangle$. Thus intuitively, the loss $\mathcal{L}_{\text{un3}}$ seeks to align the representation generated by the self-attention mechanism with the classification head vector $\mathbf{v}_k$.

6. Experiments

6.1. Setup

Based on previous studies (Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2022d), we evaluate our method on widely used vision semi-supervised benchmark datasets, including CIFAR-10/100, STL-10, and ImageNet. Our approach (OTMatch) incorporates the optimal transport loss with the calculation of the unsupervised loss within FreeMatch. Our experiments primarily focus on realistic scenarios with limited labeled data. We utilize SGD as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. The learning rate follows a cosine annealing scheduler, initialized at 0.03. The batch size is set to 64, except for ImageNet where it is 128. The ratio of unlabeled data to labeled data is 7. We report results over multiple runs over seeds. Regarding the choice of backbones, we use the Wide ResNet28-2 for CIFAR-10, Wide ResNet-28-8 for CIFAR-100, Wide ResNet-37-2 for STL-10, and ResNet-50 for ImageNet. Our training process consists of 2^{20} total training iterations, where each step involves sampling an equal number of labeled images from all classes. For the hyperparameter settings of our method, we set $\lambda = 0.5$ for CIFAR-10, $\lambda = 0.15$ for STL-10 and CIFAR-100, and

Table 2. Error rates (100% - accuracy) on ImageNet with 100 labels per class.

	Top-1	Top-5
FixMatch (Sohn et al., 2020)	43.66	21.80
FlexMatch (Zhang et al., 2021)	41.85	19.48
FreeMatch (Wang et al., 2022d)	40.57	18.77
OTMatch (Ours)	**39.29**	**17.77**

Table 3. Comparisons with state-of-the-art semi-supervised learning methods on Amazon Review and Yelp Review. Error rates (100% - accuracy) are reported.

	Amz Review		**Yelp Review**	
# Label	250	1000	250	1000
FixMatch (Sohn et al., 2020)	47.85	43.73	50.34	41.99
CoMatch (Li et al., 2021)	48.98	44.37	46.49	41.11
Dash (Xu et al., 2021)	47.79	43.52	35.10	30.51
AdaMatch (Berthelot et al., 2022)	46.75	43.50	48.16	41.71
SimMatch (Zheng et al., 2022)	47.27	43.09	46.40	41.24
FlexMatch (Zhang et al., 2021)	45.75	43.14	46.37	40.86
OTMatch (Ours)	**43.81**	**42.35**	**43.61**	**39.76**

Table 4. Ablation studies on the chosen cost in the optimal transport loss. Error rates (100% - accuracy) on CIFAR-10 with 4 labels per class are reported.

	Top-1
Binary Cost	5.20
Cost Based on Covariance	4.88
OTMatch Cost (Ours)	**4.72**

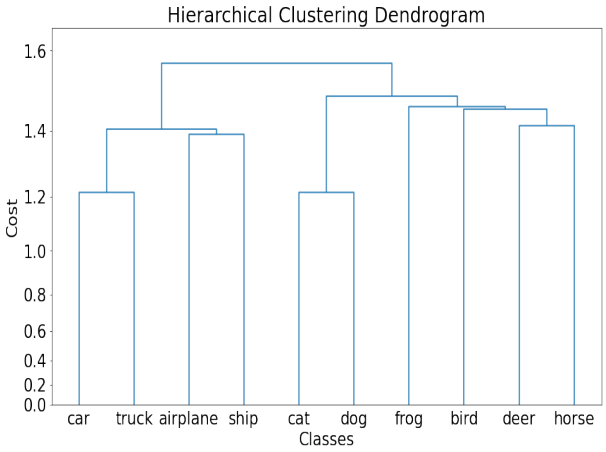

Figure 2. Hierarchical clustering results of the learned cost matrix on CIFAR-10.

$\lambda = 0.01$ for ImageNet. The momentum coefficient of the cost update is set to 0.999.

6.2. Results

Performance improvements. In our evaluation, we compare our approach to a wide range of representative semi-supervised learning methods, including Π-Model (Rasmus et al., 2015a), Pseudo-Label (Lee et al., 2013), VAT (Miyato et al., 2018), MeanTeacher (Tarvainen & Valpola, 2017), MixMatch (Berthelot et al., 2019b), ReMixMatch (Berthelot et al., 2019a), UDA (Xie et al., 2020), Dash (Xu et al., 2021), MPL (Pham et al., 2021), FixMatch (Sohn et al., 2020), FlexMatch (Zhang et al., 2021), and FreeMatch (Wang et al., 2022d). The results are reported in Table 1 and 2.

It's evident that OTMatch outperforms previous methods across the board and notably enhances performance. This improvement is particularly pronounced in scenarios with limited labeled samples such as STL-10 with 40 labels and CIFAR-10 with 10 labels, which indeed aligns with our motivation. It is important to note that in CIFAR-10 cases, fully supervised has achieved an error rate of 4.62 (Wang et al., 2022d). Thus, our method further closes the gap between semi-supervised learning and fully supervised learning.

Furthermore, as optimal transport can be incorporated wherever cross-entropy is used, our method can seamlessly integrate with recent and future techniques thereby achieving greater performance enhancements. The computational complexity is only $O(K)$, making it computationally friendly even as the number of labels scales. This highlights optimal transport as a useful regularizer with minimal computation overhead.

6.3. Analysis

Results on data from other modalities. To demonstrate the utility of our approach, we further extend our evaluations to encompass USB datasets (Wang et al., 2022c) of language modality. Specifically, the results in Table 3 demonstrate that on both Amazon Review and Yelp Review, when our approach is integrated with Flex-Match (current state-of-the-art), we have achieved an improvement, reaching a new state-of-the-art. This also validates the fact that, beyond the compatibility with FreeMatch, our approach actually can be effortlessly integrated with various existing methods.

Ablations of the cost. The cost plays a crucial role in the optimal transport loss, we conduct further ablation studies to investigate its effect. The binary cost $\mathbf{c}_1(x, y) = \mathbb{I}_{x \neq y}$ is a straightforward cost option. However, it does not take into account the relationships between classes. Additionally, we explore an alternative cost formulation that considers class relationships. In this regard, we update the cost based on the covariance matrix of predicted probabilities for strongly augmented images. We compare the performance of these costs on CIFAR-10 with 40 labels benchmark and summarize the results in Table 4. It is clear that the cost used in our method achieves the best performance.

In addition, we also demonstrate the hierarchical clustering results of the final cost matrix in Fig. 2, revealing that correct inter-class semantic relationships are learned. Specifically, In the hierarchical clustering, we can observe that various classes that denote non-living things such as airplane, truck, and ship are clustered closely together, while classes that denote living things (animals) like cat, dog, frog, bird, deer, and horse are also clustered closely together. More interestingly, we can also observe more fine-grained clustering effects, such as the proximity between car and truck, cat and dog, as well as deer and horse.

7. Conclusion

In this paper, we present a fresh perspective for semi-supervised learning, going beyond the previous efforts on solely improving the quality of pseudo-labels. We introduce a novel algorithm, OTMatch, that harnesses the inherent relationships between classes with inverse optimal transport. We also demonstrate the superiority of OTMatch in our experiments.

By introducing OTMatch, we not only contribute to the

advancement of semi-supervised learning techniques but also pave the way for future research by promoting the incorporation of optimal transport loss in a versatile manner.

Acknowledgment

Weiran Huang is supported by 2023 CCF-Baidu Open Fund and Microsoft Research Asia.

We would also like to express our sincere gratitude to the reviewers of ICML 2024 for their insightful and constructive feedback. Their valuable comments have greatly contributed to improving the quality of our work.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., and Rabbat, M. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8443–8452, 2021.

Berthelot, D., Carlini, N., Cubuk, E. D., Kurakin, A., Sohn, K., Zhang, H., and Raffel, C. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b.

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., and Kurakin, A. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *ICLR*, 2022.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021a.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021b.

Chen, H., Tao, R., Fan, Y., Wang, Y., Wang, J., Schiele, B., Xie, X., Raj, B., and Savvides, M. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *International Conference on Learning Representations (ICLR)*, 2023a.

Chen, M., Du, Y., Zhang, Y., Qian, S., and Wang, C. Semi-supervised learning with multi-head co-training. In *AAAI*, pp. 6278–6286, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020b.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chen, Y., Tan, X., Zhao, B., Chen, Z., Song, R., Liang, J., and Lu, X. Boosting semi-supervised learning by exploiting all unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7548–7557, 2023b.

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

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a wasserstein loss. *Advances in neural information processing systems*, 28, 2015.

Gong, C., Wang, D., and Liu, Q. Alphamatch: Improving consistency for semi-supervised learning with alpha-divergence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13683–13692, 2021.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Hu, Z., Yang, Z., Hu, X., and Nevatia, R. Simple: similar pseudo label exploitation for semi-supervised classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15099–15108, 2021.

Huang, Z., Shen, L., Yu, J., Han, B., and Liu, T. Flatmatch: Bridging labeled data and unlabeled data with cross-sharpness for semi-supervised learning. *Advances in Neural Information Processing Systems*, 36:18474–18494, 2023.

Kantorovich, L. On the transfer of masses (in russian). In *Doklady Akademii Nauk*, volume 37, pp. 227, 1942.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Laine, S. and Aila, T. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Lee, D.-H. 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, pp. 896, 2013.

Li, J., Xiong, C., and Hoi, S. C. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

Li, Y. J. X., Chen, Y., He, Y., Xu, Q., Yang, Z., Cao, X., and Huang, Q. Maxmatch: Semi-supervised learning with worst-case consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Nassar, I., Herath, S., Abbasnejad, E., Buntine, W., and Haffari, G. All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7241–7250, 2021.

Nassar, I., Hayat, M., Abbasnejad, E., Rezatofighi, H., and Haffari, G. Protocon: Pseudo-label refinement via online clustering and prototypical consistency for efficient semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11641–11650, 2023.

Pham, H., Dai, Z., Xie, Q., and Le, Q. V. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11557–11568, 2021.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rasmus, A., Berglund, M., Honkala, M., Valpola, H., and Raiko, T. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 28: 3546–3554, 2015a.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. Semi-supervised learning with ladder network. *ArXiv*, abs/1507.02672, 2015b.

Shi, L., Zhang, G., Zhen, H., Fan, J., and Yan, J. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. 2023.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Taherkhani, F., Dabouei, A., Soleymani, S., Dawson, J., and Nasrabadi, N. M. Transporting labels via hierarchical optimal transport for semi-supervised learning. In *European Conference on Computer Vision*, pp. 509–526. Springer, 2020.

Tai, K. S., Bailis, P. D., and Valiant, G. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. In *International Conference on Machine Learning*, pp. 10065–10075. PMLR, 2021.

Tan, Z., Wang, Z., and Zhang, Y. Seal: Simultaneous label hierarchy exploration and learning. *arXiv preprint arXiv:2304.13374*, 2023a.

Tan, Z., Yang, J., Huang, W., Yuan, Y., and Zhang, Y. Information flow in self-supervised learning. *arXiv preprint arXiv:2309.17281*, 2023b.

Tan, Z., Zhang, Y., Yang, J., and Yuan, Y. Contrastive learning is spectral clustering on similarity graph. *arXiv preprint arXiv:2303.15103*, 2023c.

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

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Lukasiewicz, T., Massiceti, D., Hu, X., Pavlovic, V., and Neophytou, A. Np-match: When neural processes meet semi-supervised learning. In *International Conference on Machine Learning*, pp. 22919–22934. PMLR, 2022a.

Wang, X., Lian, L., and Yu, S. X. Data-centric semi-supervised learning. *arXiv preprint arXiv:2110.03006*, 2021.

Wang, X., Wu, Z., Lian, L., and Yu, S. X. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14647–14657, 2022b.

Wang, Y., Chen, H., Fan, Y., SUN, W., Tao, R., Hou, W., Wang, R., Yang, L., Zhou, Z., Guo, L.-Z., Qi, H., Wu, Z., Li, Y.-F., Nakamura, S., Ye, W., Savvides, M., Raj, B., Shinozaki, T., Schiele, B., Wang, J., Xie, X., and Zhang, Y. USB: A unified semi-supervised learning benchmark for classification. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022c.

Wang, Y., Chen, H., Heng, Q., Hou, W., Savvides, M., Shinozaki, T., Raj, B., Wu, Z., and Wang, J. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022d.

Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268, 2020.

Xu, Y., Shang, L., Ye, J., Qian, Q., Li, Y.-F., Sun, B., Li, H., and Jin, R. Dash: Semi-supervised learning with dynamic thresholding. In *International Conference on Machine Learning*, pp. 11525–11536. PMLR, 2021.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

Zhang, Y., Tan, Z., Yang, J., Huang, W., and Yuan, Y. Matrix information theory for self-supervised learning. *arXiv preprint arXiv:2305.17326*, 2023a.

Zhang, Y., Yang, J., Tan, Z., and Yuan, Y. Relationmatch: Matching in-batch relationships for semi-supervised learning. *arXiv preprint arXiv:2305.10397*, 2023b.

Zheng, M., You, S., Wang, F., Qian, C., Zhang, C., Wang, X., and Xu, C. Ressl: Relational self-supervised learning with weak augmentation. *Advances in Neural Information Processing Systems*, 34:2543–2555, 2021.

Zheng, M., You, S., Huang, L., Wang, F., Qian, C., and Xu, C. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, 2022.

Appendix

A. More on proofs

A.1. Proof of lemma 4.2

Proof. By using the definition of Wasserstein distance. We find that $\mathcal{W}(\delta_x, \sum_{i=1}^m \frac{1}{m}\delta_{s_i}) = \frac{1}{m}\sum_{i=1}^m (x - s_i)^2$. As this is a quadratic function of x, we can immediately derive that the unique minimizer is $\frac{\sum_{i=1}^m s_i}{m}$. □

A.2. Proof of lemma 5.1

Proof. Note $\mu(i) \leq \nu(i)$ for any $i \neq k$, Thus by the probability constraints we know that $\mu(k) \geq \nu(k)$. As the cost is generated by a metric, we know that $\mathbf{C}_{kk} = 0$. Consider transporting mass from ν to μ, as the cost from k to k is 0, ν will transport all $\nu(k)$ to $\mu(k)$. For any $i \neq k$, if ν transport $\Delta > 0$ mass to point j ($j \neq k$). Then as $\nu(j) \geq \mu(j)$, ν can only transport the mass Δ to the unique point where ν has smaller mass than μ. From triangular inequality $\mathbf{C}_{ik} \leq \mathbf{C}_{ij} + \mathbf{C}_{jk}$, this is costly than transporting directly from i to k. Thus the optimal plan is to transport all the residual mass $\nu(i) - \mu(i)$ to node k. Thus the conclusion follows. □

B. More algorithms derived from semantic-distribution matching with optimal transport

Shi et al. (2023) show that SimCLR (Chen et al., 2020a) and MoCo (He et al., 2019) can be understood by the optimal transport viewpoint. We would like to show that many other import algorithms can also be derived from optimal transport. Other understandings of self-supervised learning can be found in (Tan et al., 2023c;b; Zhang et al., 2023a).

B.1. Cross-entropy based contrastive methods

SimMatch (Zheng et al., 2022), CoMatch (Li et al., 2021), ReSSL (Zheng et al., 2021), SwAV (Caron et al., 2020) and DINO (Caron et al., 2021a) adopt the teacher student setting and use KL divergence in their loss (consider the effect of stop-gradient, the cross-entropy is equivalent to KL divergence). Similar to the derivation in Section 4, the teacher (student) matching matrices are generated by setting the cost matrix using the (negative) similarity of query samples between buffer samples (SimMatch, ReSSL), class prototypes (SwAV), other samples in a batch (CoMatch) or classification head weights (DINO). The derivation is similar to Section 4.

While both our OTMatch and contrastive learning-based methods consider the relationship between classes, there are some crucial distinctions. Our OTMatch focuses on aligning the class classification probabilities of two augmented views, following the line of work such as FixMatch, FlexMatch, and FreeMatch. In contrast, contrastive learning-based methods emphasize the consistency between two batches of augmented views. To be more specific, our OTMatch calculates each optimal transport loss exclusively involving the two augmented views. In contrast, contrastive learning-based methods such as SimMatch and CoMatch align the two batches by utilizing the representation similarity between samples in the batch. As a result, contrastive learning-based methods necessitate an additional branch, apart from the one calculating sample-wise consistency. Therefore, our OTMatch is orthogonal to contrastive learning-based methods and can be combined with them.

B.2. CLIP

CLIP (Radford et al., 2021) is a multi-modal learning algorithm. For a batch of image text pairs $\{(I_i, T_i)\}_{i=1}^B$, the image to text loss is as follows:

$$\mathcal{L}_{I\to T} = -\sum_{i=1}^{B} \log \frac{\exp(\langle f_I(I_i), f_T(T_i)\rangle/\tau)}{\sum_{k=1}^{B} \exp(\langle f_I(I_i), f_T(T_k)\rangle/\tau)}, \tag{10}$$

where f_I is the image encoder and f_T is the text encoder and τ is the temperature.

The loss can be retrieved by setting the cost matrix as $\mathbf{C}_{i,j} = \|f_I(I_i) - f_T(T_j)\|_2^2$ and the ground truth matching matrix $\bar{\mathbf{T}} = \text{diag}\frac{1}{B}\mathbf{1}_B$. By noticing the fact that representations are normalized and using equation (8), calculating the loss in IOT will give the loss $\mathcal{L}_{I\to T}$. Different from the uni-modal case where there will only be transportation between the single modality. In multi-modal cases, there will also exist a symmetric transportation loss $\mathcal{L}_{T\to I}$, which can be explained by optimal transport similarly.

B.3. SupCon

SupCon (Khosla et al., 2020) is a supervised learning method that generates compact representations of images by incorporating label information. Suppose I a batch of augmented images and $A(i) = I - \{i\}$. Denote $\mathbf{z}_i$ as the representation of image i, its label is $\tilde{\boldsymbol{y}}_i$.

$$\mathcal{L}_{\text{SupCon}} = \sum_{i \in I} -\frac{1}{|P(i)|} \log \sum_{p \in P(i)} \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_a / \tau\right)}. \tag{11}$$

Here, $P(i) \equiv \{p \in A(i) : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$ is the set of indices of all positives in the batch.

By setting $\mathbf{C}_{ii} = +\infty$ and $\mathbf{C}_{ij} = c - \mathbf{z}_i \cdot \mathbf{z}_j$. Noticing that the i-th row of ground truth matching matrix $\bar{\mathbf{T}}_{ij} = \frac{1}{I|P(i)|} \delta_{\tilde{\boldsymbol{y}}_i}^{\tilde{\boldsymbol{y}}_j}$. By noticing the fact that representations are normalized and using equation (8), calculating the loss in IOT will give the SupCon loss.

B.4. BYOL and SimSiam

BYOL (Grill et al., 2020) and SimSiam (Chen & He, 2021) uses the MSE loss between two augmented views. For a batch of images $\{\mathbf{x}_i\}_{i=1}^B$, we usually apply different augmentations to the images and get two batches of representations $\{\mathbf{z}_i^{(1)}\}_{i=1}^B$ and $\{\mathbf{z}_i^{(2)}\}_{i=1}^B$.

Take $\mu = \frac{1}{B}\mathbf{1}_B$, Shi et al. (2023) using the optimal value of the following optimization problem 12 to explain SimCLR and MoCo. We consider first change the inner minimization of the entropic regularization problem in 12 into the common optimal transport problem and get optimization problem 13. However, this bi-level optimization problem is still hard to solve. Thus we then relax the problem into an optimization problem 14.

Take the ground-truth matching matrix as $\bar{\mathbf{T}} = \operatorname{diag} \frac{1}{B}\mathbf{1}_B$. The cost matrix $\mathbf{C}_{i,i} = \|\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\|_2^2$ and $\mathbf{C}_{i,j} = c + \|\mathbf{z}_i^{(1)} - \mathbf{z}_i^{(2)}\|_2^2$ $(j \neq i)$, where c is a large constant.

Then the optimization problem will be Const.$+\frac{1}{B}\sum_i - \log \mathbf{T}_{ii} + \lambda \sum_i (\mathbf{C}_{i,i}\mathbf{T}_{i,i} + \sum_{j \neq i}(c + \mathbf{C}_{i,i})\mathbf{T}_{i,j})$. Using the constraint of $U(\mu)$ and simplifying the constant out, the objective function will be $\frac{1}{B}\sum_i(-Bc\lambda\mathbf{T}_{i,i} - \log \mathbf{T}_{i,i}) + \lambda \sum_i \mathbf{C}_{i,i} + \frac{c\lambda}{B}$. As $\mathbf{T} \in U(\mu)$, the optimal value is Const. $+ \lambda \sum_i \mathbf{C}_{ii}$. This exactly recovers the MSE loss.

$$\begin{aligned} \min \quad & \mathrm{KL}(\bar{\mathbf{T}} \| \mathbf{T}^\theta) \\ \text{subject to} \quad & \mathbf{T}^\theta = \underset{\mathbf{T} \in U(\mu)}{\arg\min} \langle \mathbf{C}^\theta, \mathbf{T} \rangle - \epsilon \mathrm{H}(\mathbf{T}). \end{aligned} \tag{12}$$

$$\begin{aligned} \min \quad & \mathrm{KL}(\bar{\mathbf{T}} \| \mathbf{T}^\theta) \\ \text{subject to} \quad & \mathbf{T}^\theta = \underset{\mathbf{T} \in U(\mu)}{\arg\min} \langle \mathbf{C}^\theta, \mathbf{T} \rangle. \end{aligned} \tag{13}$$

$$\begin{aligned} \min \quad & \mathrm{KL}(\bar{\mathbf{T}} \| \mathbf{T}) + \lambda \langle \mathbf{C}, \mathbf{T} \rangle \\ \text{subject to} \quad & \mathbf{T} \in U(\mu). \end{aligned} \tag{14}$$

B.5. Integrating OTMatch with self-supervised learning method

We also provide a solution for applying our training algorithm to self-supervised learning methods like DINO (Caron et al., 2021b). Given probability distributions p (teacher distribution) and q (student distribution), the cross-entropy (CE) loss is defined as $-\sum p_i \log q_i$. As p uses sharpening, the proxy losses can be defined in the following way:

$$\sum_{i=1}^{K} \mathbf{C}_{ik} \mid p_i - q_i \mid, \text{where } k = \arg\max p.$$

Table 5. Accuracy (%) on CIFAR-10.

DINO	88.06
DINO+OTMatch	**88.20**

Table 6. Per-iteration running time

FreeMatch	0.23 s
OTMatch	0.25 s

where $\mathbf{C}$ represents the cost in optimal transport. The results are shown in Table 5.

C. Analysis of the running time

We calculate the per-iteration running time of FreeMatch and OTMatch on CIFAR-10 with 40 labels. From Table 6, it can be observed that our method introduces a small computation overhead.