OT-Filter: An Optimal Transport Filter for Learning with Noisy Labels

Chuanwen Feng† Yilong Ren† Xike Xie
University of Science and Technology of China
Data Darkness Lab, MIRACLE Center, Suzhou Institute for Advanced Research, USTC
{chuanwen,ylren}@mail.ustc.edu.cn, {xkxie}@ustc.edu.cn

Abstract

The success of deep learning is largely attributed to the training over clean data. However, data is often coupled with noisy labels in practice. Learning with noisy labels is challenging because the performance of the deep neural networks (DNN) drastically degenerates, due to confirmation bias caused by the network memorization over noisy labels. To alleviate that, a recent prominent direction is on sample selection, which retrieves clean data samples from noisy samples, so as to enhance the model's robustness and tolerance to noisy labels. In this paper, we revamp the sample selection from the perspective of optimal transport theory and propose a novel method, called the OT-Filter. The OT-Filter provides geometrically meaningful distances and preserves distribution patterns to measure the data discrepancy, thus alleviating the confirmation bias. Extensive experiments on benchmarks, such as Clothing1M and ANIMAL-10N, show that the performance of the OT-Filter outperforms its counterparts. Meanwhile, results on benchmarks with synthetic labels, such as CIFAR-10/100, show the superiority of the OT-Filter in handling data labels of high noise.

1. Introduction

Deep learning has achieved great success on a flurry of emerging applications, such as [28, 29, 32, 49]. It is believed that the phenomenal achievement of deep learning is largely attributed to accurate labels. However, the inaccuracy or imprecision of labels is inherent in real-world datasets, the so-called *noisy label challenge*. One way to alleviate that is to collect labels from internet queries over data-level tags, but the performance of deep neural networks (DNN) suffers drastically from the inaccuracy of such labels [27,32]. A higher quality of data labels can be obtained by employing human workers, but the seemingly "ground truth annotations" inevitably involve human biases or mistakes [45,66]. More, the human annotation is expensive and time-consuming, especially for large-scale datasets.

There have been many works on handling noisy labels, such as regularization [23] and transition matrix [13, 43]. The regularization approach leverages the regularization bias to overcome the label noise issue. But the regularization bias is permanent [62], thus overfitting models to noisy labels. The transition matrix approach assumes that the transition probabilities between clean and noisy labels are fixed and independent of data samples. However, a quality label transition matrix is hard to be estimated, especially when the number of classes is big, making it fall short in handling noisy real-world datasets, such as [60] and [36]. A recent prominent direction is on adopting sample selection [27, 38, 58] for enhancing the label quality, by selecting clean samples from the noisy training dataset. In general, existing literatures in sample selection can be grouped to two categories, co-training networking [27, 62] and criterion-based filtering [31, 34, 57, 58].

The former utilizes the memorization of DNNs and multiple networks (e.g., co-teaching [27] and its variants [38, 62]) to filter label noise with small loss trick, so that a small set of clean samples are used as training examples. Letting alone the high training overhead of multiple networks, the disadvantages are two-fold: 1) it may require the a-prior knowledge of noisy rates to select the specified proportion of small loss samples as clean samples; 2) the small-loss trick is not tolerant to the error accumulation of network training once a clean sample is falsely recognized, the so-called confirmation bias, especially for labels with high noise where clean and noisy samples largely overlap.

The latter alleviates the problem by setting a specific criterion. Mostly, existing works [31] [58] adopt Euclidean distances for measuring the similarity between data samples. Distance-based filtering iteratively explores the neighborhood of the feature representation and infers/cleans sample labels by aggregating the information from their neighborhoods. Despite the simplicity, distance-based filtering is insufficient to address noisy labels, especially when the label noise is high, e.g., overlapped label classes.

†Equal Contribution.

In this paper, we revamp the sample selection from the perspective of optimal transport [56] and propose a novel filtering method, called the *OT-Filter*. In light of the optimal transport, we construct discrete probability measures over the sample features, which lifts the Euclidean space of feature vectors to a probability space. It thus enables a geometric way of measuring the discrepancy between probability measures, representing the corresponding sample features. In addition to the distance-based metric in the Euclidean space [31, 58], the OT-Filter also captures the distribution information in the probability space so as to alleviate the confirmation bias. Accordingly, a clean representation can be obtained for each class in the probability space. By optimizing the transport plan from a sample to a clean representation, one can better determine if a sample is clean or noisy, thus improving the quality of sample selection.

In general, the merits of the OT-Filter can be summarized as follows. First, it does not require any a-prior knowledge about the noisy rate of a dataset. Second, it utilizes the optimal transport which provides geometrically meaningful distances to exploit the sample discrepancy yet preserving the distribution patterns in corresponding probability space, making the sample selection of high quality and theoretical support. Third, it can be plugged to existing robust training paradigms, e.g., supervised and semi-supervised robust training.

We conduct extensive experiments with a series of synthetic and real datasets to gain insights into our proposals. The result shows that the OT-Filter follows state-of-the-art (SOTA) [38] when the noise rate is low, and dominates SOTA when the noise rate is high. For instance, the OT-Filter achieves about 14% and 12% higher accuracy than SOTA, in the presence of 90% noise rate, on synthetic datasets CIFAR-10 and CIFAR-100, respectively. Moreover, the OT-Filter outperforms the competitors on real datasets Clothing1M and ANIMAL-10N.

The rest of the paper is organized as follows. We review the existing literature in Section 2. In Section 3, we present preliminaries of optimal transport. We investigate our proposed OT-Filter in Section 4. Furthermore, we conduct extensive empirical studies in Section 5 and conclude the paper in Section 6.

2. Related Work

2.1. Learning with Noisy Label

A flurry of research methods [5, 13, 14, 27, 40, 54] was proposed for learning with noisy labels, which can be roughly divided into two categories, robust model [13, 22, 39, 43, 54, 61, 64] and clean sample selection [27, 32, 62]. The clean sample selection aims to select a clean subset from a noisy dataset [14, 27, 38]. One way is to train multiple networks to filter noisy samples. For instance, [32] proposed to pre-train an extra network to select clean samples for the main network. The inferiority of this method is on error accumulation. To relieve that, Han et al. [27] proposed Co-teaching that maintains two networks simultaneously, where one network selects clean samples for the other network by small loss trick. However, as the increase of training epochs, Co-teaching would converge to consensus gradually. Then, Co-teaching$^+$ [62] was proposed to keep two networks diverged. To avoid confirmation bias, DivMix [38] proposed a hybrid framework with semi-supervised training, achieving the state-of-the-art performance. Based on DivMix, [41, 65] modified the training schema, such as self-supervised training [11, 33], to boost the performance.

The other research line of sample selection is to filter noisy samples based on specific criteria [31, 34, 42, 59]. For example, Xia et al. [59] proposed to reduce the uncertainty of loss using specific strategies. Iscen et al. [31] proposed to leverage similarities between training examples in the feature space, encouraging the prediction of each example to be close to its nearest neighbors. Also, in Euclidean space, Wu et al. [58] proposed the TopoFilter that filters noisy samples using k-nearest neighborhood distance between prelogits. Fine [34] proposed to use the principal components of latent representations to select clean samples.

2.2. Optimal Transport

The optimal transport problem [55, 56] aims to move mass from a probability measure to another probability measure at a minimum cost. It defines the Wasserstein distance [24] and provides a geometric way to measure the discrepancy between probability measures. One factor that limits the wide application of optimal transport is the high computational cost. To improve its scalability, Cuturi [17] proposed an entropic regularization for the transport plan, which yields an efficient algorithm with the matrix scaling method of Sinkhorn-Knopp [35]. Recently, computational optimal transport [44, 48] has found many applications in various areas, e.g., generative models [4, 25], domain adaptation [15, 16], and semi-supervised learning [52, 53]. In particular, [1] studied the barycenter in Wasserstein space. [18] explored the efficient computing of Wasserstein barycenter. [19] proposed to use entropic optimal transport loss, based on joint distribution optimal transport [15], to build a robust training.

In this paper, we study the problem of clean sample selection from the perspective of optimal transport to combat label noise. The proposed method can potentially capture geometric information from the probability space. Unlike the mechanism of multiple network training, it only trains a single network without a-prior information on noise rates. Moreover, it can be easily plugged to multiple robust training paradigms.

3. Preliminary

In this section, we present preliminaries on optimal transport, making a basis for subsequent proposed techniques. The optimal transport [56] is to seek an optimal transport plan between two measures at a minimal cost (e.g. the Wasserstein distance), which provides a geometric way of matching probability measures. Here we first introduce the general definition of Wasserstein distance induced by optimal transport problem. More details about optimal transport can be found in [44, 48, 55].

Wassertein Distance. Let $\mathcal{S}$ be a locally complete and separable metric space, the $\mathcal{P}(S)$ be the Borel probability measures set on $\mathcal{S}$. For any $\mathcal{X}, \mathcal{Y} \subset \mathcal{S}$, given $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$, the optimal transport defines a Wasserstein distance between two probability measures, denoted as

$$W_p(\mu, \nu) := \left(\inf_{\pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^p d\pi(\mu,\nu) \right)^{\frac{1}{p}}$$

$p \geq 1$, where the $\pi(\mu, \nu)$ is the set of joint probability measures with marginal μ and ν.

3.1. Discrete Optimal Transport

Let $\Delta_n = \{\mathbf{a} \in \mathbb{R}_+^n | \sum_{i=1}^n a_i = 1, \forall a_i \geq 0\}$ be a probability simplex in dimension n. Consider two empirical probability measures $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, defined on metric space $\mathcal{X}$ with support $\{x_i\}_{i=1}^n$ and $\mathcal{Y}$ with support $\{y_j\}_{j=1}^m$ respectively. Here the weight vector $\mathbf{a} = (a_1, a_2, ..., a_n)$ and $\mathbf{b} = (b_1, b_2, ..., b_m)$ live in Δ_n and Δ_m, respectively. The δ stands for the Dirac unit mass function. Given a transport cost $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the discrete optimal transport between probability measures μ and ν can be formulated as

$$\begin{aligned} W_p^p(\mu, \nu) &:= \min_{M \in \Pi(\mu,\nu)} \langle C, M \rangle_F \\ s.t.\ M\mathbf{1}_m &= \mu,\ M^T\mathbf{1}_n = \nu \end{aligned} \tag{1}$$

where $C \in \mathbb{R}_+^{n\times m}$ is the transport cost matrix, and c_{ij} represents a unit transport cost from x_i to y_j. The $M \in \mathbb{R}_+^{n\times m}$ is transport plan in which m_{ij} denotes the amount of mass transported from x_i to y_j. All feasible transport plans constitute transport polytope $\Pi(\mu, \nu)$. The $\langle C, M \rangle_F$ is the Frobenius inner product of matrices and equals to $tr(C^T M)$.

3.2. Regularized Optimal Transport

The discrete optimal transport formulation, in essence, is a convex optimization problem. More precisely, it's a linear programming problem. Unfortunately, this linear programming problem has a cubic computing complexity. A way to relieve this is to leverage entropic regularization [17] formulated as

$$\min_{M \in \Pi(\mu,\nu)} \langle C, M \rangle_F - \epsilon H(M) \tag{2}$$

where $\epsilon > 0$ is the regularization coefficient, and $H(M)$ is the entropic regularization term, which promotes an efficient computation for transportation via the matrix scaling algorithm [17], given by

$$H(M) := -\sum_{ij} M_{ij}\big(log(M_{ij}) - 1\big)$$

3.3. Wasserstein Barycenter

Given N probability measures $\{\nu_1, \nu_2, ..., \nu_N\}$, $\nu_i \in \mathcal{P}(\mathcal{S})$, each of which has finite supports and second moments. A Wasserstein barycenter of these measures is a probability measure μ, satisfying:

$$\mu := \inf \sum_{i=1}^N \lambda_i W_2^2(\nu_i, \mu), \quad s.t. \sum_{i=1}^N \lambda_i = 1,\ \forall \lambda_i \geq 0$$

The notion was first proposed by [1], where some elegant properties of Wasserstein barycenter were presented. Then [2] discussed it on the discrete case and showed that the problem of finding Wasserstein barycenter over the space $\mathcal{P}(\mathcal{S})$ can be reduced to a simpler space $\mathcal{O}_r(S)$, where $r = \sum_{i=1}^N e_i - N + 1$, and e_i is the number of components of $\nu_i, i \in [1, N]$. To find Wasserstein barycenter over space $\mathcal{O}_r(S)$, a set of efficient algorithms [7, 18, 26] were proposed.

4. Methods

4.1. Overview

In this section, we study the OT-Filter, the proposed method based on optimal transport for learning with noisy labels in a semi-supervised paradigm. The key idea of our proposed method is to transport samples with noisy labels to samples with clean labels at a minimum cost, which makes sample selection of high quality and theoretical support.

The mechanism of the OT-Filter, consists of two phases, *representation phase* (Section 4.2) and *transportation phase* (Section 4.3). Then, the OT-Filter can be plugged to off-the-shelf SSLs for robust training, e.g., MixMatch [8], by treating clean samples as labeled data and noisy samples as unlabeled samples (Section 4.4).

4.2. Representation Phase: Finding Clean Representations via Wasserstein Barycenters

In the representation phase, we first use a pre-trained network, e.g., ResNet [29], to extract a set of feature representations of the labeled samples. Then, we define a discrete probability measure over the feature representations of each class to find the clean representations (or prototypes) via Wasserstein barycenters. Since there are mislabeled samples, the barycenters we obtained may be noisy. Therefore, we iteratively optimize the barycenters with the expectation

maximization (EM) algorithm [20] to converge to clean representations. The details about the representation phase are covered in Algorithm 1.

Feature Extraction. The purpose of the feature extraction is to transform the input noisy label dataset into a feature-label dataset via a network. A noisy label dataset of N samples can be represented by $\hat{\mathcal{D}} = \{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^d$ denotes the i-th sample and $\hat{y}_i \in \{0,1\}^K$ denotes the corresponding noisy label over K classes. Then, a feature extraction network $f(\omega, \mathbf{x}_i) : \mathbb{R}^d \to \mathbb{R}^m$ with parameters ω maps a sample $\mathbf{x}_i$ to a m-dimensional feature representation, denoted as $\tilde{\mathbf{x}}_i \in \mathbb{R}^m$. Therefore, all output feature representations and their corresponding labels constitute a feature-label set, denoted as $\tilde{D} = \{(\tilde{\mathbf{x}}_i, \hat{y}_i)\}_{i=1}^N$.

Probability Measure Modeling. The purpose of probability measure modeling is to define a probability measure over the feature-label set. This definition will potentially lift a Euclidean space $\mathcal{D}$, in which the feature representations originally live, to a probability space, abbreviated as $\mathcal{P}(\mathcal{D})$. Assume that the labels are of K classes. First, we split the feature-label set $\tilde{D}$ into K subsets, i.e, $\{s_1, s_2, ..., s_K\}$ according to the label class. For one feature representation $\tilde{\mathbf{x}}_i$ in class $k \in [1, K]$, we represent it as $\tilde{\mathbf{x}}_i^k, i \in [1, |s_k|]$. Then, all feature representations in subset s_k construct a probability measure, denoted as

$$\mathcal{Q} = \frac{1}{|s_k|} \sum_{i=1}^{|s_k|} \delta_{\tilde{\mathbf{x}}_i^k}$$

where the $\delta_{\tilde{\mathbf{x}}_i^k}$ is a Dirac unit mass on $\tilde{\mathbf{x}}_i^k$, and for simplicity, we use uniform weights.

Clean Representation Retrieval. The problem of finding the representation of any subset s_k is equivalent to finding the corresponding Wasserstein barycenter $\mathcal{B}_k$ with finite supports, formalized as follows:

$$\mathcal{B}_k := \inf_{\mathcal{B} \in \mathcal{O}_r(\tilde{\mathcal{D}})} W_2^2(\mathcal{B}, \mathcal{Q}) \tag{3}$$

where the W is the Wasserstein distance, and $r = |s_k|$. As mentioned in Section 3, from the perspective of linear programming, the above optimization admits the dual:

$$D(\alpha, \beta) = \max_{(\alpha,\beta) \in R(C)} \alpha^T \mathcal{B} + \beta^T \mathcal{Q}$$

where the polyhedron of dual variables is:

$$R(C) = \{(\alpha, \beta) \in \mathbb{R}^r \times \mathbb{R}^r | \alpha_i + \beta_j \leq C_{ij} \wedge i, j \in [r]\}$$

By means of duality theory [6,9,10], we can find the relation of solution between the primal and dual problem. Here, the optimal solution of $D(\alpha, \beta)$ is a subgradient of $W_2^2(\mathcal{Q}, \mathcal{B})$ with respect to $\mathcal{B}$. Thus, we employ a simple projected subgradient [18] to optimize Equation 3. over $\mathcal{B}$.

Despite we have found a set of Wasserstein barycenters $\{\mathcal{B}_i\}_{i=1}^K$, the barycenters may be noisy since there are mislabeled data. Moreover, the clean barycenters are unknown. Therefore, for each class, we consider the ideal and clean barycenter $\mathcal{B}_c$ as a hidden variable, the obtained noisy barycenter $\mathcal{B}$ as an observed variable. Then, the problem of finding clean barycenter $\mathcal{B}_c$ is to maximize the following log-likelihood:

$$\hat{\theta} \leftarrow \arg\max_{\theta} \log \sum_{\mathcal{B}_c} p(\mathcal{B}_c, \mathcal{B} | \theta)$$

where the θ is the parameters of probabilistic model. By this way, we can optimize the noisy barycenters iteratively with the EM algorithm to converge to clean barycenters. In E-step, we construct optimal transport between all samples and clean barycenters, thereby we can infer the labels of all samples. In M-step, we refine the barycenters with inferred label obtained from E-step. The overall computing details see Algorithm 1.

Algorithm 1 Find Clean Representation via Wasserstein Barycenter

Require: $\mathcal{Q} = \{\tilde{\mathbf{x}}_i\}_{i=1}^{|s_k|} \in \mathbb{R}^{m \times r}$, regularization coefficient ϵ, $\mathcal{B} \in \mathbb{R}^{m \times e}$

Initialize: $\mathcal{B}$, $\nu \leftarrow \mathcal{B}$, $\mu \leftarrow \mathcal{Q}$, $\eta = 1/2$, $t = 2$
while $\mathcal{B}$ not converged **do**
  set $\nu = \mathbf{1}_e/e$, $\mu = \mathbf{1}_r/r$, $\hat{\nu} = \tilde{\nu} = \mathbf{1}/e$
  while ν not converged **do**
    $\beta \leftarrow (t+1)/2$, $\nu \leftarrow (1-\beta^{-1})\hat{\nu} + \beta^{-1}\tilde{\nu}$
    $\alpha \leftarrow$ optimizer of $\mathbf{d}(\nu, \mu, C)$
    $\tilde{\nu} \leftarrow \tilde{\nu} \circ \exp(-\beta\alpha)$, $\tilde{\nu} \leftarrow \tilde{\nu}/\tilde{\nu}^T\mathbf{1}$
    $\hat{\nu} \leftarrow (1-\beta^{-1})\hat{\nu} + \beta^{-1}\tilde{\nu}$
    $t \leftarrow t+1$
  end while
  $\nu \leftarrow \hat{\nu}$
  $\mathcal{M} \leftarrow$ optimizer of $\mathbf{p}(\nu, \mu, C)$
  $\mathcal{B} \leftarrow (1-\eta)\mathcal{B} + \eta(\mathcal{Q}\mathcal{M})diag(\nu^{-1})$
end while
while $\mathcal{B}_c$ not converged **do**
  E-Step:
  $\mathbb{L}_{\hat{\mathcal{D}}} \leftarrow$ OT(μ, ν, C)
  M-Step:
  $\mathcal{B}_c \leftarrow$ refine $\mathcal{B}$ with $\mathbb{L}_{\hat{\mathcal{D}}}$
end while
Return $\mathcal{B}_c$

4.3. Transportation Phase: Transporting Noisy Labels via Regularized Optimal Transport

In the transportation phase, noisy feature representations are transported to clean representations. In particular, we align all feature representations, including those of noisy

samples, to the clean representations (obtained in the representation phase) via sparsity regularized optimal transport. This operation potentially detects the corrupted labels based on the result of the optimal transport. Here, the mechanism of our sample selection via optimal transport is described as follows. First, we align the clean barycenters to the all rest samples via optimal transport. Based on the result of optimal transport, we can infer the labels of the all rest samples. We consider the route with the max mass transport that forms a coupling and then assign the label of a clean barycenter to the samples on this route. Second, we select the sample whose inferred label is identical to the original label as a clean sample.

Sparsity Regularization. After the representation phase, clean representations are obtained, which are essentially the barycenters of label classes. The clean representation of a class provides the guidance of identifying whether samples are clean or noisy w.r.t. the given class.

We consider the barycenters of all the K classes as a discrete probability measure, denoted as $\mu = \frac{1}{K}\sum_{k=1}^{K}\delta_{u_k}$, where δ_{μ_k} is a Dirac unit mass on μ_k. Accordingly, the feature representations of the entire dataset construct a discrete probability measure $\nu = \frac{1}{N}\sum_{i=1}^{N}\delta_{\tilde{\mathbf{x}}_i}$. Then, we transport feature representations ν of all samples to the clean barycenters μ, in order to detect if a sample is clean or mislabeled. Inspired by [16, 46], we view the clean barycenters as the source domain and the feature representations of samples as the target domain. Then, we recast the alignments as an optimal transport problem with a sparsity regularization [16], written as

$$\min_{M\in\Pi(\mu,\nu)} \langle C, M\rangle_F - \epsilon H(M) + \gamma \sum_{j,c} ||M(i_c, j)||_1^{\frac{1}{2}} \quad (4)$$

The second term is an entropic regularization, as described in Equation 2. The third term is a sparsity regularization with coefficient γ, where i_c index the line if its element in class c, and $M(i_c, j)$ is a vector consisting of the j-th column of M in class c. The $||\cdot||_1^{1/2}$ denotes an L_1 norm with the power of $\frac{1}{2}$. The entropic regularization enables discrete optimal transport to be efficient. However, it disperses the transport route, which negatively affects sample selection. Therefore, to improve the quality of sample selection, we employ a sparsity regularization. The sparsity regularization promotes a feature representation that would be aligned to one of the clean barycenters and penalizes transportation matrix M that aligns together feature representations with different labels.

Optimization. Despite the consistent sparsity achieved by the regularization, the objective function of Equation 4 is non-convex. Moreover, the L_1 norm with the power in label regularization term is a concave function. A common optimization strategy [37] is to construct a convex upper bound, represented by convex functions, over the non-convex problem. By doing so, we can use the optimal solution of convex upper bound to approximate the optimal solution of the original non-convex problem. If applying the simplest linear approximation, we have

$$\sum_{j,c} ||M(i_c, j)||_1^{\frac{1}{2}} \leq \langle G, M\rangle_F + \mathcal{L}$$

where matrix G is $\frac{1}{2}(||\hat{M}(i_c, j) + \delta||^{-\frac{1}{2}})$, and $\hat{M}$ is a given start point. The steps of optimization process are depicted as Algorithm 2.

Algorithm 2 Transportation: Transport Noisy Label via Regularized Optimal Transport

Require: optimal cost C_{min} from Equation 1,
1: Initialize: $G = \mathbf{0}$
2: **while** G not converged **do**
3: $C_{min} \leftarrow C_{min} + G$
4: $M \leftarrow$ optimizer of Equation 2 with C_{min}
5: $G \leftarrow$ update G using M
6: **end while**

4.4. Robust Training in SSL

After all feature representations are transported to the clean barycenters. We filter noisy samples and obtain two data subsets, namely the clean sample set and noisy sample set. As aforementioned, our sample selection method is flexible enough for supporting various robust training paradigms, e.g., the robust supervised training [64] and the robust semi-supervised learning [50]. Following [38], we select MixMatch [8] with data augmentation strategy [41] as the practice of robust semi-supervised training but with fewer operations. We admit the labels of clean samples and neglect the labels of noisy samples, and then we use labeled data and unlabeled data for training in a semi-supervised learning fashion. The overall training process is described as Algorithm 3.

5. Experiments

5.1. Datasets

We perform extensive experiments on four benchmark datasets: CIFAR-10 [36] and CIFAR-100 [36] with synthetic label noise, and Clothing1M [60] and ANIMAL-10N [51] with real label noise.

CIFAR10/100. We evaluate the OT-Filter on CIFAR-10 and CIFAR-100 [36] with synthetic noise. Both datasets contain $60,000$ 32×32 color images, in 10 and 100 classes, respectively. From both datasets, 50K images are used for training, and 10K images are used for testing. Since noise characteristics can hardly be determined in advance, synthetic noise is commonly taken for controlling the noise rate

Algorithm 3 Robust Training in SSL

Require: network parameter ω, Beta distribution parameter α, weight of unlabeled loss $\lambda_{\mathcal{U}}$, batch size B, number of augmentations K, sharpening temperature T, labeled data $\mathcal{X}$, unlabeled data $\mathcal{U}$

while b in B **do**
  $\hat{x}_b \leftarrow \text{augment}(x_b)$
  for k in K **do**
    $\hat{u}_{b,k} \leftarrow \text{augment}(u_b)$
  end for
  $\hat{q}_b \leftarrow \frac{1}{K}\sum_k p_{model}(\hat{u}_{b,k};\omega)$
  $q_b \leftarrow \text{sharpen}(\hat{q}_b, T)$
end while
$\hat{\mathcal{X}} = \{(\hat{x}_b, p_b)\}_{b=1}^{B}, \hat{\mathcal{U}} = \{(\hat{u}_{b,k}, q_b)\}_{b,k=1}^{B,K}$
$\mathcal{W} = \text{shuffle}\ (concat(\hat{\mathcal{X}}, \hat{\mathcal{U}}))$
$\hat{\mathcal{X}}^{'} = Mixup(\hat{\mathcal{X}}_i, \mathcal{W}_i), i \in [1, |\hat{\mathcal{X}}|]$
$\hat{\mathcal{U}}^{'} = Mixup(\hat{\mathcal{U}}_i, \mathcal{W}_{i+\hat{\mathcal{U}}}), i \in [1, |\hat{\mathcal{U}}|]$
$\mathcal{L}_{\mathcal{X}} \leftarrow \text{CE}(\mathcal{X}^{'}), \mathcal{L}_{\mathcal{U}} \leftarrow \text{MAE}(\mathcal{U}^{'})$
$\mathcal{L} \leftarrow \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}}\mathcal{L}_{\mathcal{U}}$

to deliberately evaluate learning algorithms with noisy labels. Thus, following previous works [27, 38], we consider two types of label noise, i.e., symmetric and asymmetric label noise. The symmetric label noise is generated by randomly flipping labels of a portion of samples from one class to all other possible classes. The asymmetric label noise is designed to follow the structure of real-world label noise, where labels are flipped to similar classes within the superclasses.

Clothing1M. Clothing1M is a large-scale dataset with noisy labels [60], containing over 1M images obtained from online shopping websites. The labels are from 14 classes generated based on surrounding texts provided by the sellers, and the noise rate is estimated around 38.5% [39]. Also, the dataset provides 3 clean datasets for training, validation, and testing, containing 50K, 14K, and 10K images, respectively.

ANIMAL-10N. ANIMAL-10N is a real-world noisy dataset released by [51], which is crawled from several online search engines using predefined labels as searching keywords. There are in total 55K images, of which 50K images are for training, and 5K images are for testing. The noise rate was estimated to be around 8%.

5.2. Implementation Details

For CIFAR-10 and CIFAR-100, we use the PreAct ResNet18 [29] as the backbone and train it using the SGD optimizer with the following settings: a momentum of 0.9, a weight decay of 5e-4, and a batch size of 128. The learning rate was initialized as 0.02 and reduced by a factor of 10, after 150 epochs. The network was trained for 300 epochs. The warmup period is 10 epochs for CIFAR-10, and 30 epochs for CIFAR-100.

For Clothing1M, we use the ResNet50 [28] pre-trained on ImageNet [21] as the backbone and train it using the SGD optimizer with the following settings: a momentum of 0.9, a weight decay of 1e-3, and a batch size of 32. The learning rate was initialized as 0.002 and reduced by a factor of 10 after 50 epochs. The network was trained for 120 epochs. For each epoch, we sample 1000 mini-batches from the training data.

For ANIMAL-10N, we use VGG19 [49] with batch normalization [30] as the backbone and train it using the SGD optimizer. We train the network for 100 epochs. The initial learning rate is set as 0.01 and reduced by a factor of 5 after 50 and 75 epochs.

In addition, we set the entropic regularization coefficient ϵ and sparsity regularization coefficient γ as 10 and 1, respectively, which are consistent for all training implementations.

5.3. Experimental Results

In this section, we present experimental results of the OT-Filter on benchmark datasets with both synthetic and real label noises. For CIFAR-10 and CIFAR-100, we consider the noise rate of 20%, 50%, 80%, and 90% for symmetric noise, and noise rate of 40% for asymmetric noise, respectively. The results regarding precision and recall were drawn from the 100-th epoch of robust training.

Sample Selection Performance. We first present the empirical study on the quality of sample selection. The dataset CIFAR-10 under noise rate of 90% (symmetric) and noise rate (asymmetric) of 40% was employed to compare our method with state-of-the-art sample selection method DivMix [38]. Here, we consider two metrics, precision and recall. As shown in Figure 1, with the increase of training epochs, our method outperforms the DivMix [38] in terms of both precision and recall.

For example, on 90% symmetric noise, when the training epoch is above 100, the precision of the OT-Filter is steadily above 90%, while the precision of DivMix is less than 73%. In terms of the recall, the OT-Filter is above 90%, while the DivMix is below 82%, when the training epochs are higher than 100. It also shows that our method converges steadily to the optimal solution, which DivMix yields fluctuations in the convergence process. Similar trends can be observed on 40% asymmetric noise.

CIFAR10 and CIFAR100 Datasets. Table 1 shows the test accuracy on CIFAR-10 for symmetric and asymmetric noise, respectively. In all testings, the OT-Filter demonstrates good performance in test accuracy. In particular, in the presence of high noise, the OT-Filter significantly outperforms other competitors. For 90% noise rate, the test accuracy of our OT-Filter is 90.5%, which is about 14%

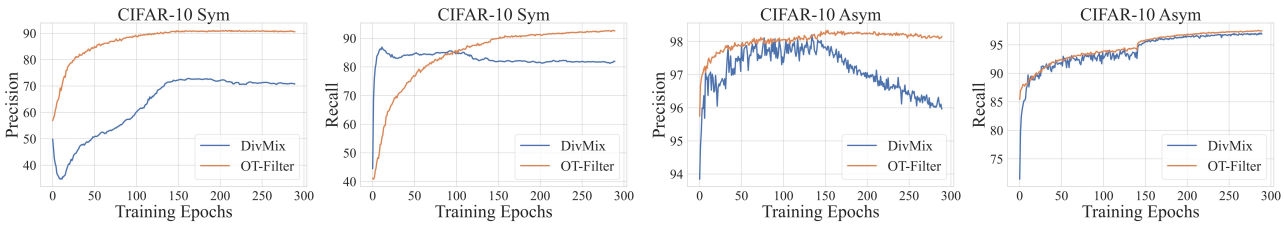

Figure 1. Sample selection performance (Precision & Recall) w.r.t. Training Epochs on CIFAR-10 under synthetic label noise.

Method	CIFAR-10 20%	50%	80%	90%	Asym 40%
CE	86.8	79.4	62.9	42.7	77.3
Co-teaching$^+$ [62]	89.5	85.7	67.4	47.9	71.3
Co-learning [54]	92.2	84.5	61.2	-	81.4
TopoFilter [58]	90.2	-	45.7	-	87.9
CRUST [40]	91.1	86.3	58.3	-	88.8
Fine [34]	91.0	87.3	69.4	-	89.5
M-correction [3]	93.6	91.8	75.8	74.7	93.3
CTRR [61]	93.1	-	83.7	81.7	89.0
DivMix [38]	**96.1**	94.6	93.2	76.0	93.4
Fine+DivMix [34]	**96.1**	94.9	93.5	**90.5**	93.8
OT-Filter	96.0	**95.3**	**94.0**	**90.5**	**95.1**

Table 1. Test accuracies(%) obtained from state-of-the-art sample selection methods. The best results are in bold. The data are copied from respective papers. The - denotes the lack of respective data.

Method	CIFAR-100 20%	50%	80%	90%	Asym 40%
CE	62.0	46.7	19.9	10.1	44.5
Co-teaching$^+$ [62]	65.6	51.8	27.9	13.7	-
Co-learning [54]	66.6	54.5	35.5	-	47.6
CRUST [40]	65.2	56.4	-	-	53.0
M-correction [3]	73.9	66.1	41.6	24.3	47.4
CTRR [61]	70.1	-	43.7	-	54.5
TopoFilter [58]	65.6	-	20.7	-	-
Fine [34]	70.3	64.2	25.6	-	61.7
DivMix [38]	77.3	**74.6**	60.2	31.5	55.1
Fine+DivMix [34]	**79.1**	**74.6**	61.0	34.3	-
OT-Filter	76.7	73.8	**61.8**	**42.8**	**76.5**

Table 2. Test accuracies(%) obtained from state-of-the-art sample selection methods. The best results are in bold. The data are copied from respective papers. The - denotes the lack of respective data.

better than DivMix. Although [38] performs slightly better than the OT-Filter at the 20% noise rate, our method is also competitive. Similar trends can be observed on CIFAR-100 (Table 2), where our method mostly follows the SOTA method and significantly outperforms its competitors when the noise rate is high.

Method	Backbone	Test Accuracy
CE	ResNet-50	69.2
M-correction [1]	ResNet-50	71.0
Co-teaching [27]	ResNet-50	71.7
CTRR [61]	ResNet-50	72.7
Fine [34]	ResNet-50	72.9
CRUST [40]	ResNet-50	73.5
TopoFilter [58]	ResNet-50	74.1
DivMix* [38]	ResNet-50	74.3
Fine+DivMix [34]	ResNet-50	74.4
OT-Filter	ResNet-50	**74.5**

Table 3. Test accuracy(%) on Clothing1M. The * denotes we have run the algorithm based on the official implementation. The data are copied from respective papers. The best result is in bold.

Clothing1M Dataset. Table 3 shows the result of test accuracy on Clothing1M by comparing different competitors. The result validates the effectiveness of the OT-Filter on the large-scale real-world dataset. Moreover, It can be observed that the OT-Filter follows the state-of-the-art performance regarding test accuracy.

Method	Test Accuracy
CE	79.4
Nested [47]	81.3
SELFIE [51]	81.8
PLC [63]	83.4
Co-teaching+Nested [12]	84.1
GJS [22]	84.2
OT-Filter	**85.5**

Table 4. Test accuracy(%) on ANIMAL-10N. The data are copied from respective papers. The best result is in bold.

ANIMAL-10N Dataset. Table 4 shows the experimental results on ANIMAL-10N, a real-world dataset with moderate label noise. We compared our method with multiple state-of-the-art baseline methods that are not limited to sample selection. For fair comparison, we use the same VGG-

Dataset	CIFAR-100							
Noise Rate	80%		90%		80%		90%	
Method	Best	Last	Best	Last	Precision	Recall	Precision	Recall
OT-Filter w/o Spar-Reg	51.0	50.7	23.8	23.7	90.4	47.1	62.1	19.7
OT-Filter	**61.8**	**61.5**	**42.8**	**42.1**	**93.7**	**61.8**	**72.3**	**40.0**

Table 5. Ablation study for sparsity regularization on CIFAR-100 dataset. The left part is the performance comparison and the right part is the quality of sample selection.

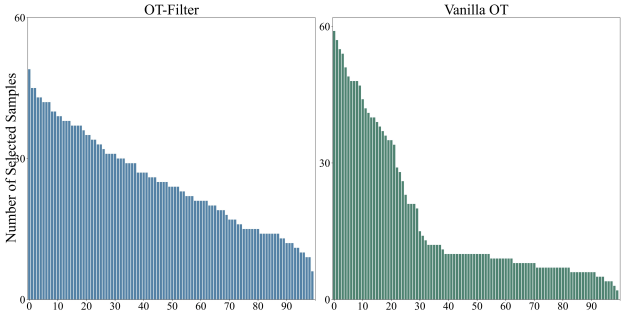

Figure 2. The ablation study for the quality of sample selection w/o sparsity regularization on CIFAR-100 under noise rate of 90%. The result shows that sparsity regularization brings a relatively uniform sample selection.

19 network architecture. The results show we outperform all competitors, and have 2.1% performance improvement over PLC [63], and 3.7% performance improvement over SELFIE [51].

5.4. Analysis

We conduct a series of ablation studies to understand the effectiveness of the key components of the OT-Filter, including the sparsity regularization of the optimal transport, the EM algorithm, and the entropic regularization coefficient.

Effect of Sparsity Regularization. The sparsity regularization discussed in Section 4.3 is one of the key components in our OT-Filter, which is designed to improve the accuracy of the optimal transport by controlling the sparsity of transport. We therefore study the effect of the sparsity regularization over sample selection and performance under high noise on CIFAR-100 in Table 5. It can be observed, with the sparsity regularization, both the quality of sample selection and test accuracy are much better than the counterpart without it. In particular, when the noise rate is 90%, it shows that the technique of sparsity regularization brings in over 10% improvement in precision, about 21% improvement in recall, and 19% improvement in performance. Moreover, as shown in Figure 2, with the sparsity regularization, the number of samples selected for each class is more uniform. It is mainly attributed to discrete optimal transport allowing mass split.

Dataset	CIFAR-10							
Noise Rate	20%		50%		80%		90%	
Method	Best	Last	Best	Last	Best	Last	Best	Last
OT-Filter w/o EM	95.9	95.7	95.0	94.8	93.6	93.4	90.0	89.8
OT-Filter	**96.0**	**95.8**	**95.3**	**95.1**	**93.7**	**93.5**	**90.5**	**90.2**

Table 6. Ablation study for EM algorithm on CIFAR-10 dataset.

Dataset	CIFAR-10							
Noise Rate	20%		50%		80%		90%	
Method	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
$\epsilon = 1$	99.7	93.9	99.0	93.9	94.6	90.6	88.6	84.2
$\epsilon = 5$	99.7	93.8	99.0	93.8	94.6	90.6	88.6	84.2
$\epsilon = 10$	99.7	93.8	99.0	93.8	94.5	90.6	88.6	84.2

Table 7. Ablation study for entropic regularization coefficient on CIFAR-10 dataset.

Effect of Expectation Maximization. To filter noisy labels, we align all samples to clean barycenters via optimal transport. Therefore, the quality of clean barycenters affects the quality of sample selection and therefore the performance of the robust training. Table 6 indicates the impact of EM on CIFAR-10 dataset. It shows that the EM helps in improving the performance under different noise rates, and the significance is higher for high noise rates.

Effect of Entropic Regularization Coefficient ϵ. The motivation for equipping the optimal transport with entropic regularization is to speed up its computation. However, the entropic regularization could potentially degrade the transport sparsity and therefore the quality of sample selection. To alleviate that, we propose sparsity regularization, whose effectiveness is examined in Table 5. Then, a by-product of sparsity regularization is the parameter robustness of the regularization coefficient ϵ. As shown in Table 7, the performance of sample selection stays quite stable when varying ϵ from 1 to 10.

6. Conclusion

In this work, we study the problem of sample selection for learning with noisy labels. We propose the OT-Filter, a novel technique for retrieving clean samples, which can also be combined with existing semi-supervised learning techniques for handling noisy labels. Unlike previous works relying on a-prior knowledge or conforming to confirmation bias, the OT-Filter enhances the measurement of data discrepancy, by lifting data from Euclidean space to the probability space, and thus improves the quality of sample selection. Extensive experiments on synthetic and real-world noisy datasets are conducted to evaluate our proposals. The results show its effectiveness and superiority.

Acknowledgement. This work is supported by NSFC (No.61772492, 62072428) and the CAS Pioneer Hundred Talents Program. Xike Xie is the corresponding author.

References

[1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. 2, 3, 7

[2] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84(2):389–409, 2016. 3

[3] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 7

[4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[5] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021. 2

[6] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear programming: theory and algorithms*. John Wiley & Sons, 2013. 4

[7] 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. 3

[8] 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. 3, 5

[9] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. 4

[10] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 4

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[12] Yingyi Chen, Xi Shen, Shell Xu Hu, and Johan AK Suykens. Boosting co-teaching with compression regularization for label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2688–2692, 2021. 7

[13] De Cheng, Tongliang Liu, Yixiong Ning, Nannan Wang, Bo Han, Gang Niu, Xinbo Gao, and Masashi Sugiyama. Instance-dependent label-noise learning with manifold-regularized transition matrix estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16630–16639, 2022. 1, 2

[14] Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021. 2

[15] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[16] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014. 2, 5

[17] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2, 3

[18] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014. 2, 3, 4

[19] Bharath Bhushan Damodaran, Rémi Flamary, Vivien Seguy, and Nicolas Courty. An entropic optimal transport loss for learning deep neural networks under label noise in remote sensing images. *Comput. Vis. Image Underst.*, 191:102863, 2018. 2

[20] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 4

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[22] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021. 2, 7

[23] Kilian Fatras, Bharath Bhushan Damodaran, Sylvain Lobry, Remi Flamary, Devis Tuia, and Nicolas Courty. Wasserstein adversarial regularization for learning with label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[24] Alessio Figalli and Federico Glaudo. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. 2021. 2

[25] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018. 2

[26] Wenshuo Guo, Nhat Ho, and Michael Jordan. Fast algorithms for computational optimal transport and wasserstein barycenter. In *International Conference on Artificial Intelligence and Statistics*, pages 2088–2097. PMLR, 2020. 3

[27] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 1, 2, 6, 7

[28] 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. 1, 6

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 1, 3, 6

[30] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 6

[31] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4672–4681, 2022. 1, 2

[32] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 1, 2

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2

[34] Taehyeon Kim, Jongwoo Ko, JinHwan Choi, Se-Young Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021. 1, 2, 7

[35] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 2

[36] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1, 5

[37] Kenneth Lange. *MM optimization algorithms*. SIAM, 2016. 5

[38] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 1, 2, 5, 6, 7

[39] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 6

[40] Baharan Mirzasoleiman, Kaidi Cao, and Jure Leskovec. Coresets for robust training of deep neural networks against noisy labels. *Advances in Neural Information Processing Systems*, 33:11465–11477, 2020. 2, 7

[41] Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8022–8031, 2021. 2, 5

[42] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021. 2

[43] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017. 1, 2

[44] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2, 3

[45] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of machine learning research*, 11(4), 2010. 1

[46] Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR, 2019. 5

[47] Oren Rippel, Michael Gelbart, and Ryan Adams. Learning ordered representations with nested dropout. In *International Conference on Machine Learning*, pages 1746–1754. PMLR, 2014. 7

[48] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. 2, 3

[49] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 6

[50] 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. 5

[51] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR, 2019. 5, 6, 7, 8

[52] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Transporting labels via hierarchical optimal transport for semi-supervised learning. In *European Conference on Computer Vision*, pages 509–526. Springer, 2020. 2

[53] Fariborz Taherkhani, Ali Dabouei, Sobhan Soleymani, Jeremy Dawson, and Nasser M Nasrabadi. Self-supervised wasserstein pseudo-labeling for semi-supervised image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12267–12277, 2021. 2

[54] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1405–1413, 2021. 2, 7

[55] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009. 2, 3

[56] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021. 2, 3

[57] Yikai Wang, Xinwei Sun, and Yanwei Fu. Scalable penalized regression for noise detection in learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 346–355, 2022. 1

[58] Pengxiang Wu, Songzhu Zheng, Mayank Goswami, Dimitris Metaxas, and Chao Chen. A topological filter for learning with label noise. *Advances in neural information processing systems*, 33:21382–21393, 2020. 1, 2, 7

[59] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. In *International Conference on Learning Representations*, 2022. 2

[60] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 1, 5, 6

[61] Li Yi, Sheng Liu, Qi She, A Ian McLeod, and Boyu Wang. On learning contrastive representations for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16682–16691, 2022. 2, 7

[62] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019. 1, 2, 7

[63] Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *Ninth International Conference on Learning Representations*, volume 9, 2021. 7, 8

[64] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 2, 5

[65] Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1657–1667, 2022. 2

[66] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017. 1