Semi-discrete Optimal Transport for Long-tailed Classification

Lianbao Jin
Dalian University of Technology
1099630577@qq.com

Na Lei
Dalian University of Technology
http://conformalgeometry.org/~lei/

Zhongxuan Luo
Dalian University of Technology
zxluo@dlut.edu.cn

Jin Wu
Huawei Technologies Co., Ltd
lion.wujin@huawei.com

Chao Ai
Huawei Technologies Co., Ltd
aichao@huawei.com

Xianfeng Gu
Stony Brook University
https://www3.cs.stonybrook.edu/~gu/

Abstract

The long-tailed data distribution poses an enormous challenge for training neural networks in classification. The classification network can be decoupled into a feature extractor and a classifier. This paper takes a semi-discrete optimal transport perspective to analyze the long-tailed classification problem, where the feature space is viewed as a continuous source domain, and the classifier weights are viewed as a discrete target domain. The classifier is indeed to find a cell decomposition of the feature space with each cell corresponding to one class. The imbalanced training set causes the more frequent classes to have larger volume cells, which means that the classifier's decision boundary is biased towards less frequent classes, resulting in reduced classification performance in the inference phase. Therefore, we propose a novel OT-dynamic softmax loss, which dynamically adjusts the decision boundary in the training phase to avoid overfitting in the tail classes. In addition, our method incorporates the supervised contrastive loss so that the feature space satisfies the uniform distribution condition. Extensive and comprehensive experiments demonstrate that our method achieves state-of-the-art performance on multiple long-tailed recognition benchmarks, including CIFAR-LT, ImageNet-LT, iNaturalist 2018, and Places-LT.

Keywords: ***Semi-discrete Optimal Transport, Long-tailed Classification, Decision Boundary, Supervised Contrastive Loss.***

1. Introduction

In recent years, the rapid development of computer vision technologies [16, 24] is inseparable from large-scale, high-quality and balanced datasets such as ImageNet [34] and MS-COCO [27]. Unlike the computer vision datasets with a roughly uniform distribution of tags, the real-world datasets always have a skewed distribution with a long tail: a few classes have a large proportion in the datasets, while most have a small proportion. Deep learning methods perform poorly on such imbalanced datasets [3, 14, 46] because neural networks often favor the majority classes.

A huge range of approaches has been proposed to solve the long-tailed classification problem, albeit in different ways, including re-sampling [3, 14, 18, 37], Balanced Softmax [17, 29, 33, 51], long-tailed contrastive learning [8, 20, 36, 42], etc [9, 21, 40]. The Balanced Softmax approaches correct the decision boundaries based on the class frequency. However, such approaches are label-aware and can be fragile when training online models in real-time since the number of samples per class is unknown or changes dynamically for different batches. In addition, the OTLM [31] uses the sinkhorn algorithm to do post-hoc correction for long-tailed learning. However, the OTLM method requires a large number of samples for evaluation. This is because if the batch size is small, the OTLM can not guarantee that the desired marginal distribution can be satisfied within the batch. Sinkhorn algorithm is theoretically calculating the optimal transport plan. Its solution is not unique, and the solution may be locally optimal. We introduce semi-discrete optimal transport into the long-tailed classification to solve the above problems.

Recently, optimal transport has become a popular tool for machine learning [1, 11, 25, 30, 39]. Essentially semi-discrete optimal transport provides a cell decomposition of the source domain such that each cell is mapped to one target point, and the measure of the cell is equal to the measure of the target point. Intuitively the classification problem is also a cell decomposition problem. A classification network

is generally composed of a feature extractor and a classifier. The feature extractor maps high-dimensional data to low-dimensional feature space. Moreover, the classifier decomposes the feature space into cells so that the same class data is within the same cell. Each class can be viewed as one target point, and the frequency of the class can be viewed as the discrete measure of the target point. Thus each cell is mapped to one corresponding class, and the measures of the cells are equal to the frequencies of the classes. The cell boundaries are the decision boundaries. Thus, the classification task can be viewed as semi-discrete optimal transport, with the feature space as the source domain and the weights of the classifier as the target points.

However, it is improper to use the frequencies as the target measures for long-tailed data. When we use a conventional training framework, minimizing the empirical cross-entropy loss will allocate more giant cells of the feature space to more frequent classes, which implies that the decision boundary is biased towards less frequent classes. In other words, the size of each cell follows the sampling frequency; more samples form a larger cell. The test set is usually balanced, which means the lower frequency classes need bigger cells in the inference process. Thus, the conventional trained network is more generalized for frequent classes, whereas its performance is lacking for infrequent classes. The imbalanced classification learning mainly depends on how the appropriate decision boundary is drawn. Semi-discrete Optimal transport can adjust the decision boundary by altering the measure of the target point. The literature [12] provides a geometric variational method to solve the semi-discrete optimal transport problem, in which the source measure is a continuous uniform distribution, and the target measure is discretized into finite points with the Dirac measure. Inspired by this, we propose a novel Optimal Transport Dynamic Softmax Loss (OT-dynamic) for long-tailed classification, which dynamically adjusts the decision boundary in the training process. Moreover, in order for the source domain to satisfy the uniform distribution as much as possible, our method uses supervised contrastive loss (SCL) [22], which has been proved to optimize for uniformity asymptotically [43].

To summarize, the main contributions of our works are three-folds:

• We propose that the classifier essentially computes the semi-discrete optimal transport problem, and the weights of the classifier are the Wasserstein centers of each class. From the perspective of semi-discrete optimal transport, we explain that the reduced accuracy of long-tailed classification is due to the offset of decision boundary toward the tail classes.

• We develop a simple and effective OT-dynamic Softmax Loss for the network training phase, which can shift decision boundary dynamically to avoid overfitting in the tail classes.

• Extensive experiments show that our method achieves more significant improvement than the current SOTA methods on several benchmarks, including the artificially imbalanced datasets CIFAR-LT [23], ImageNet-LT [34] and the natural world large scale imbalanced datasets iNaturalist2018 [41], Places-LT [56].

2. Related Work

2.1. Long-tailed classification

Re-sampling and Balanced Softmax are the most intuitive approaches to deal with long-tailed classification. Re-sampling strategies can be further divided into two types: Under-sampling the head classes [3, 14, 37] and Over-sampling the tail classes [4, 18, 54]. Balanced Softmax [17, 29, 33, 51] is another outstanding strategy. It corrects the decision boundaries based on the class frequency. However, such approaches are label-aware and easily overfitted in the tail classes since, during online training models in real-time, the number of samples per class is unknown or changes dynamically for different batches. Compared with Balanced Softmax, our approach can effectively avoid overfitting the tail classes. Other approaches have also been proposed to solve the long-tailed classification problem. For instance, the literature [21] introduced to decouple the learning phase into representation learning and classifier fine-tuning. The papers [44, 48, 55] proposed the multi-expert structure for the long-tailed problem.

2.2. Contrastive learning

Recently, contrastive learning has shown great promise in unsupervised representation learning [6, 15]. SimCLR [6] is the first to match the performance of a supervised ResNet with only a linear classifier trained on self-supervised representation on large-scale datasets. MoCo [15] uses a momentum encoder to maintain a consistent representation of negative pairs extracted from the memory library. Supervised comparative learning [22] is an extension of comparative learning, which can obtain better feature representation by incorporating the label information to compose positive and negative pairs. The paper [43] proved that the contrastive loss optimizes for alignment and uniformity asymptotically. Many researchers are also exploring long-tailed contrastive recognition. The paper [20] proposed k-positive contrast loss to learn a balanced feature space to reduce class imbalance and improve model generalization. After that, [36] introduced typical contrastive learning into hybrid networks to enhance long-tail learning. DRO-LT [42] extended the prototype contrastive learning optimization with distributed robustness, which makes the learning model more robust. Paco [8] further innovated supervised comparative learning by adding a set of parameter

learnable class centers. In this paper, we illustrate that the classifier weights are the Wasserstein centers of each class from the perspective of semi-discrete optimal transport.

2.3. Optimal transport

Optimal transport has been widely used in machine learning, such as the following applications: generative model [1, 10, 13, 25, 35], domain adaption [7, 50], 3D shape matching [38], graph matching [49], and model designs [19]. In particular, the literature [31] proposed the OTLM method to do post-hoc correction for long-tailed learning. Our work proposes OT-dynamic softmax loss from the perspective of semi-discrete optimal transport, which is used in the training phase.

3. Method

This section presents the theoretic analysis for the long-tailed classification from a semi-discrete optimal transport perspective and introduces our computational algorithm.

3.1. Semi-discrete optimal transport problem

In this subsection, we will introduce basic concepts in classic optimal transport theory, focusing on the solution of semi-discrete optimal transport. We refer readers to [12, 32] for detailed derivation.

The optimal transport problem is to find a map that minimizes the cost of interdomain transport and ensures the quantity of measurement. Suppose X, Y are two subsets of an m-dimensional Euclidean space with probability measures μ and ν, respectively. We require μ and ν share the same total measure, i.e.,

$$\int_X d\mu = \int_Y d\nu = 1 \tag{1}$$

A map $T : X \to Y$ is measure-preserving if for any measurable set $B \subset Y$, the set $T^{-1}(B)$ is μ-measurable and

$$\mu(T^{-1}(B)) = \nu(B) \tag{2}$$

The measure-preserving map can also be written as $T_{\#}\mu = \nu$, where $T_{\#}\mu$ is the push-forwarded measure induced by T. Given a cost function $c(x, y) : X \times Y \to R$, which represents the cost of moving a unit mass from x to y. The total transport cost of the map T is defined as :

$$\int_X c(x, T(x))d(\mu(x)) \tag{3}$$

Monge's optimal transport problem is to find the measure-preserving map that minimizes the total transport cost,

$$\mathcal{W}_c^2(\mu, \nu) = \min_{T_{\#}\mu=\nu} \int_X c(x, T(x))d\mu(x) \tag{4}$$

The solution to Monge's problem is called the optimal transport map $\widetilde{T}$, whose total transport cost is the square of the Wasserstein distance between μ and ν, which is denoted as $\mathcal{W}_c(\mu, \nu)$.

For quadratic Euclidean distance cost, Brenier [2] proved the existence, uniqueness and the intrinsic structure of the optimal transport mapping.

Theorem 1 [2]. *Suppose X and Y are the Euclidean space $\mathbb{R}^m$ and the transport cost is the quadratic Euclidean distance cost $c(x, y) = 1/2\|x - y\|^2$. Furthermore μ is absolutely continuous and μ and ν have finite second order moments, $\int_X |x|^2 d\mu(x) + \int_Y |y|^2 d\nu(y) < \infty$, then there exists a convex function $u : X \to \mathbb{R}$, the so-called Briener potential, its gradient map ∇u gives the solution to the Monge's problem,*

$$(\nabla u)_{\#}\mu = \nu \tag{5}$$

The Brenier potential is unique upto a constant, hence the optimal transport mapping is unique.

Brenier's theorem can be directly generalized to the discrete situation. Suppose the source measure μ (uniform distribution) has a continuous density $\mu \in L^1(\Omega)$, defined on a convex domain $\Omega \subset \mathbb{R}^m$, and the target measure is discrete $\nu = \sum_{k=1}^K \nu_k \delta_{w_k}$, $\nu_k > 0$, $w_k \in \Omega$, μ and ν share the same total measure $\int_\Omega d\mu = \sum_{k=1}^K \nu_k = 1$, then the Monge's problem becomes the semi-discrete optimal transport problem, which has a nice geometric characterization.

As shown in Fig. 1, each target point $\boldsymbol{w_k}$ corresponds to a supporting hyperplane of the Brenier potential as follow:

$$\pi_{h,k}(\boldsymbol{x}) = \langle \boldsymbol{w_k}, \boldsymbol{x} \rangle + h_k \tag{6}$$

Theorem 1 claims that the semi-discrete optimal transport map is given by the gradient map of Brenier potential. The Brenier potential $u_h : \Omega \to \mathbb{R}$ is a piecewise linear convex function, which is the upper envelope of all supporting hyperplanes $u_h(\boldsymbol{x}) := max_{k=1}^K \{\pi_{h,k}(\boldsymbol{x})\}$.

The graph is Brenier potential is a convex polytope. Each facet of the polytone corresponds to a suppoting hyperplane $\pi_{h,k}(\boldsymbol{x})$. The projection of the polytope induces a cell decomposition of Ω, and each cell $W_k(\boldsymbol{h})$ is the projection of the supporting plane $\pi_{h,k}(\boldsymbol{x})$. The semi-discrete optimal transport mapping induces a cell decomposition of $\Omega = \bigcup_{k=1}^K W_k$ such that each cell W_k is mapped to the corresponding target point $\boldsymbol{w_k}$ and the μ-volume of the cell W_k equals to the discrete measure ν_k of point $\boldsymbol{w_k}$.

The height vector $\boldsymbol{h}$ is the only parameter to be optimized. Obviously the larger h_k induces a larger μ-volume of the cell W_k. According to [12], $\boldsymbol{h}$ is the minimal point of the following convex energy under the condition that $\sum_k h_k = 0$,

$$E(h) = \int_0^h \sum_{k=1}^K \omega_k(\eta) d\eta_k - \sum_{k=1}^K h_k \nu_k \tag{7}$$

where $\omega_k(\eta)$ is the μ-volume of $W_k(\eta)$. The convex en-

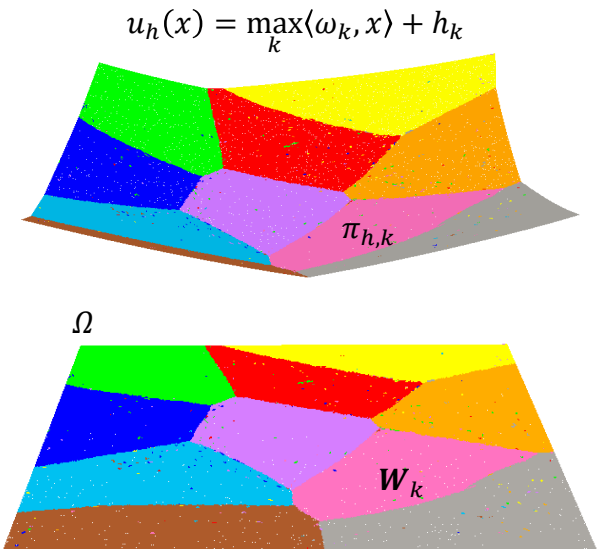

Figure 1. Brenier potential and the corresponding power diagram.

ergy $E(\boldsymbol{h})$ can be optimized directly by the gradient descend method with $\nabla E(h_k) = (\omega_k(\boldsymbol{h}) - \nu_k)$.

The key problem is to calculate the μ-volume $\omega_k(h)$ of the cell $W_k(\eta)$, which can be estimated using the Monte-Carlo method. We draw N random samples $\{\boldsymbol{x_i}\}$ from μ distribution. For each sample $\boldsymbol{x_i}$, we can find the cell W_k containing it by:

$$k = \arg\max_k \{\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + h_k\} \tag{8}$$

The μ-volume of W_k is estimated as $\hat{\omega}_k(\boldsymbol{h}) := \#\{i \mid x_i \in W_k(\boldsymbol{h})\}/N$, which converges to $\omega_k(\boldsymbol{h})$, when N goes to infinity. Accordingly, the gradient of the energy is approximated as $\nabla E(\boldsymbol{h}) \approx (\hat{\omega}_k(\boldsymbol{h}) - \nu_k)^T$. Thus we can minimize the energy by the Gradient Descent algorithm.

3.2. Optimal transport view of the classification problem

Consider a K-classification deep model $\mathcal{M}$ with a training set $\mathcal{S} = \{(\boldsymbol{s_i}, \boldsymbol{y_i})\}_{i=1}^n$, where $\boldsymbol{s_i}$ denotes i-th training sample and $\boldsymbol{y_i}$ denotes its corresponding one-hot label over K classes. Our goal is to learn the model parameters from training datasets so that $\mathcal{M}$ achieves optimal performance on evaluation datasets. We typically decompose a deep network model $\mathcal{M}$ into two components: a feature extractor f_θ and a classifier. The feature extractor f_θ first extracts a feature representation $\boldsymbol{x}_i$,

$$\boldsymbol{x}_i = f_\theta(\boldsymbol{s}_i) \in \mathbb{R}^m \tag{9}$$

which is then fed into the classifier to compute class prediction scores π_{ik} and the classifier predicts the class label j as follows:

$$\begin{aligned} \pi_{ik} &= \langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + h_k \\ k &= \arg\max_k \{\pi_{ik}\} \end{aligned} \tag{10}$$

where $\boldsymbol{w}_k$ is the k-th vector of the classifier's weight and h_k is the k-th component of the bias.

Compare Eqn. 8 and Eqn. 10 (the two formulas are the same), it is clear that the classification problem is essentially a semi-discrete optimal transport problem. As shown in Fig. 1, the feature extractor f_θ map the samples from ambient space onto the feature space $\Omega \subset \mathbb{R}^m$, which is continuous source domain. From the perspective of optimal transport, the classifier is a family of hyperplanes in the Euclidean space, which is equivalent to Brenier potential function. The weight $\boldsymbol{w}_k$ is the discrete target point and the bias h_k corresponds to the Dirac measures ν_k. The classifier task means we need to find a cell decomposition of the feature space and each cell will be mapped to one class with the least cost. Samples with the same category should be projected to the same cell, and the boundary of the cell is defined as the decision boundary. Here the weight $\boldsymbol{w}_k$ represents the k-th class so it should be the center of the samples belonging to the k-th class. By the way, The frequency of the k-th class gives the measure of the target point $\boldsymbol{w}_k$ and determines the volume of the cell. More explanation is given in the next subsection.

The training process is to learn the classifier weights $\boldsymbol{w}_k$'s and the bias h_k's. Since $\boldsymbol{w}_k$ represents the k-th class, so the total distance in the feature space between all samples $\boldsymbol{x}_i$ belonging to the k-th class and $\boldsymbol{w}_k$ should be minimal, i.e.,

$$\boldsymbol{w}_k = \arg\min_{w \in \Omega} \left\{ \sum_{y_{ik}=1} c(\boldsymbol{x_i}, \boldsymbol{w}) \right\}, k = 1, \cdots, K \tag{11}$$

where $c(\boldsymbol{x_i}, \boldsymbol{w})$ represents the transport cost from $\boldsymbol{x}_i$ to $\boldsymbol{w}$, which means the weight $\boldsymbol{w}_k$ is exactly the Wasserstein center of all the samples belonging to the k-th class.

In a word, the training process of a classification task is first embedding the samples into the feature space and then calculating the Wasserstein centers of each class represented as the classifier's weights. At last, it computes a semi-discrete optimal transport map to get the bias according to the measure of each class.

3.3. Semi-discrete optimal transport for long-tailed classification

Traditionally the measure of each class is given by the number of samples belonging to this class, which is based on the hypothesis that all classes are sampled uniformly. But for the long-tailed training set, the hypothesis does not hold anymore. According to Sec. 3.2, the measure of target

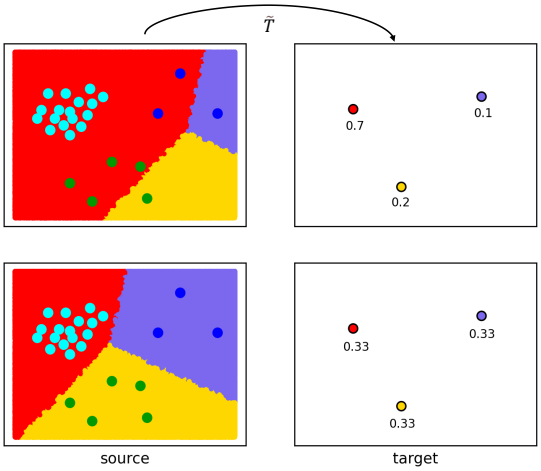

Figure 2. The continuous source domain and discrete targets are with different distributions. The classification task is equivalent to the optimal transport map $\widetilde{T}$ from source domian to targets . Different probability distributions of the targets induce different cell decompositions of the source domain. The decision boundary can be adjusted by changing the probability distribution of the targets to improve the classification performance.

is related to the frequency of classes. The more frequent classes are, the bigger measures of targets are and the larger cells are formed. That is to say, the imbalanced training set causes the decision boundary of the classifier to be biased towards less frequent classes. The key issue is the decision boundary, which is difficult to be assigned due to the imbalanced distribution of the different classes. Optimal transport can adjust the decision boundary only by modulating the bias $\boldsymbol{h}$.

Fig. 2 explains this insight. For example, in a 3-classification problem, given a set of training samples, the feature extractor maps them onto the feature space. Each sample is color-encoded to indicate its class. The light blue colored class has the highest frequency and the dark blue colored class has the lowest frequency. The classifier decomposes the feature space into three cells, the red cell for the light blue class, the yellow cell for the green class, and the purple cell for the dark blue class. As shown in the top row of Fig. 2, the light blue samples have the highest frequency; hence its corresponding red cell has the largest volume. So many green and dark blue samples are misclassified. In other words, the imbalance of the data makes the decision boundary biased to the less frequent classes (yellow and dark blue), causing the classifier to over-fit the head class (light blue) and under-fit the tail class (dark blue and yellow). From the bottom row of Fig. 2, we can see that when increasing the measure of the target, the decision boundary will move towards the class with more frequency; that is, the cell volume of the class with less frequency will also become larger.

OT-dynamic softmax loss. The above discussion shows that the performance of the classifier will be poor in the balanced test set if we directly train the model with the imbalanced training set using the following common cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log \frac{e^{\pi_{ik}}}{\sum_{j=1}^{K} e^{\pi_{ij}}} \tag{12}$$

In previous works [10-11], common logit adjustment methods subtract a positive adjusting term from π_{ij} to form an adjusted logit. Thus we get the formulation of the balanced softmax loss:

$$\mathcal{L}_{BS} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log \frac{n_k e^{\pi_{ik}}}{\sum_{j=1}^{K} n_j e^{\pi_{ij}}} \tag{13}$$

where n_k is the number of samples in the k-th class of the training set and y_{ij} is the j-th component of the one-hot label of the sample s_i. However, the Balanced Softmax is label-aware and can be fragile when training online models in real-time, since the number of samples per class is unknown or varies dynamically across batches. Therefore, we propose OT-dynamic softmax loss, which dynamically adjusts the decision boundary during training.

According to Sec. 3.1, the volume of W_k can be estimated as $\hat{\omega}_k(\boldsymbol{h}) := \#\{i \mid x_i \in W_k(\boldsymbol{h})\}/N$ in each batch, where N is the batch size, the $\#$ symbol represents the number of all samples belonging to the W_k. As shown in Fig.1, the volume of the cell W_k is uniquely determined by the bias $\boldsymbol{h}_k$. The higher the bias $\boldsymbol{h}_k$ is , the larger the cell W_k is. The gradient $\nabla E(\boldsymbol{h}_k) = (\hat{\omega}_k(\boldsymbol{h}) - \nu_k)$ of the bias $\boldsymbol{h}$ provides a direction to adjust the decision boundary, where ν_k is set to n_k/n. Specifically, when $\nabla E(\boldsymbol{h}_k) > 0$, it means that too many samples fall in W_k and the volume of W_k needs to be reduced, that is, the bias $\boldsymbol{h}_k$ needs to be reduced and the decision boundary moves to the larger cell. So in the training process, we plus $\nabla E(\boldsymbol{h}_k)$ in Alg. 1 to the bias $\boldsymbol{h}$ to dynamically adjust the decision boundary,

$$\pi_{ik} - \nabla E(\boldsymbol{h}_k) = \langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle + h_k - \nabla E(\boldsymbol{h}_k) \tag{14}$$

Furthermore, $\nabla E(\boldsymbol{h}_k)$ is dynamic, an OT-dynamic softmax function can be obtained:

$$\mathcal{L}_{OTD} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log \frac{n_k e^{\pi_{ik} - \nabla E(\boldsymbol{h}_k)}}{\sum_{j=1}^{K} n_j e^{\pi_{ij} - \nabla E(\boldsymbol{h}_j)}} \tag{15}$$

where $\nabla E(\boldsymbol{h}_j)$ is the j-th component of $\nabla E(\boldsymbol{h})$. The key idea is to adjust the bias by adding a perturbation term $\nabla E(\boldsymbol{h}_j)$ to change the volume of the corresponding cell, and further deform the decision boundary. Experiments show that compared with balanced softmax, our method can avoid overfitting in the tail class.

Supervised contrastive loss. Khosla et al. [22] extended the unsupervised contrastive loss [6] with lable information into supervised contrastive loss. The key difference between supervised contrastive loss and unsupervised contrastive loss lies in the composition of the positive and negative samples of an anchor sample. For unsupervised contrastive loss, the positive sample is only an alternatively augmented view of the anchor sample. For supervised contrastive loss, apart from the alternatively augmented counterpart, the positives also include some other samples from the same class.

In Alg. 1, the volume $\hat{\omega}_k(\boldsymbol{h})$ is estimated using the Monte Carlo method, which requires that the feature representation $\boldsymbol{x_i}$ satisfy uniform distribution. Therefore, our method adds supervised contrastive loss as follow:

$$\mathcal{L}_{SCL} = \sum_{i=1}^{N} -\frac{1}{|\{\boldsymbol{x_i^+}\}|} \sum_{\boldsymbol{x_j} \in \{\boldsymbol{x_i^+}\}} log \frac{e^{(\boldsymbol{x_i} \cdot \boldsymbol{x_j}/\tau)}}{\sum_{\boldsymbol{x_k}, k \neq i} e^{(\boldsymbol{x_i} \cdot \boldsymbol{x_k}/\tau)}} \quad (16)$$

Here, $\{\boldsymbol{x_i^+}\} = \{\boldsymbol{x_j} | y_j = y_i, i \neq j\}$ is the set of all positives in the different views batch distinct from anchor $\boldsymbol{x_i}$, and $|\{\boldsymbol{x_i^+}\}|$ is its cardinality. The $\cdot$ symbol denotes the inner product, $\tau > 0$ is a scalar temperature parameter.

Algorithm 1 OT-dynamic Softmax Loss

Input: the logit of neural network $\hat{\mathbf{y}}$, the label $\mathbf{y}$
Output: OT-dynamic Softmax Loss

1: Calculate $\hat{\omega}_k(\boldsymbol{h}) = \#\{i | \arg\max_k(\hat{\mathbf{y}}_i) \in W_k(\boldsymbol{h})\}/N$
2: Calculate $\nabla E(\boldsymbol{h}) = (\hat{\omega_k}(\boldsymbol{h}) - \nu_k)^T$
3: $\nabla E(\boldsymbol{h}) = \nabla E(\boldsymbol{h}) - mean(\nabla E(\boldsymbol{h}))$
4: $\hat{\mathbf{y}} \leftarrow \hat{\mathbf{y}} - \nabla E(\boldsymbol{h})$
5: **return** $Balanced\ Softmax(\hat{\mathbf{y}}, \mathbf{y})$

Overview of our method. In practice, we train the classification network with a combination of two losses. An optimal transport dynamic softmax loss is applied to the output of the classification layer to adjust the decision boundary dynamically, and the supervised contrastive loss is applied to the feature representation of the penultimate layer to satisfy the uniform distribution. Thus our method linearly combines these two losses:

$$\mathcal{L} = \lambda \mathcal{L}_{OTD} + (1 - \lambda)\mathcal{L}_{SCL} \quad (17)$$

where λ is a loss weight hyper-parameter and is set to 0.5 in all experiments.

4. Experiments

In this section, we conduct a series of experiments to evaluate the effectiveness of our algorithm using ResNet/ResNeXt as the baseline architecture. However, our method and analysis are not limited to those architectures. All experimental setup is the same as that [17]. Below we first introduce the long-tailed benchmarks and experimental setting details in Sec. 4.1 and Sec. 4.2, respectively. Then, we report the experimental results compared with revelant methods on long-tailed classification task in Sec. 4.3. Finally, we conduct ablation study to show that OT-dynamic softmax loss can avoid overfitting in the tail classes and improve the robustness of the training process in Sec. 4.4.

4.1. Datasets Details

CIFAR-10/100-LT. CIFAR-10 and CIFAR-100 [23] have $50,000$ images for training and $10,000$ images for validation with ten categories and 100 categories, respectively. Following the prior work [5, 55], we use the imbalance factors ρ to control the degree of data imbalance degrees. $\rho = \frac{N_{max}}{N_{min}}$, where N_{min} and N_{max} are the number of training samples for the least frequent and the most frequent classes respectively. We conduct experiments with ρ equals to 10, 100 and 200 respectively.

ImageNet-LT. ImageNet-LT is a subset by sampling from ImageNet-2012 [34] following the Pareto distribution with the power value $\alpha = 6$. It contains $115.8K$ images for training, $20K$ for validation and $50K$ for testing. The overall number of categories is 1000, with the number of images per class ranging from 5 to 1280.

iNaturalist 2018. The iNaturalist 2018 [41] has $437,513$ training images from 8124 classes, with an imbalance factor of 500. It is a real-world dataset for species, large scale and extremely imbalanced. In addition, the iNaturalist 2018 also face the fine-grained problem [47].

Places-LT. Places-LT [56] is a long-tailed version of the large-scale scene classification dataset Places. It consists of $184.5K$ images from 365 categories with class cardinality ranging from 5 to $4,980$. We use the same training and validation splits strategy for fair comparisons as [28] in our experiments.

4.2. Experimental Setting

Implementation Details. For the long-tailed CIFAR datasets, we adopt ResNet-32 as our backbone network and SGD optimizer with the momentum 0.9 for all experiments. The learning rate increased from 0.05 to 0.1 in the first 800 iterations. Through all the experiments, the batch size is 256, and the model is end-to-end trained for $13K$ iterations. The experimental setup is the same as that in [33].

For a fair comparison, we use the same experimental setting as [21] on ImageNet-LT, Places-LT, and iNaturalist 2018. For Places-LT, following the works in [33, 21], we choose ResNet-152 as the backbone network and pretrain it on the full ImageNet-2012 dataset.

On ImageNet-LT, we use ResNet-10, ResNet-50 and ResNeXt-50, respectively. On iNaturalist 2018, we use

ResNet-50. For ResNet-10 and ResNeXt-50, we adopt cosine learning rate schedule gradually decaying from 0.05 to 0, with image resolution 224×224 and batch size 128. For the ResNet-152, we use a batch size of 128 for the limited GPU memory. For all experiments, we adopt SGD optimizer with momentum 0.9 on 4 NVIDIA 1080Ti GPUs.

Evaluation Protocol. After training on the long-tailed datasets, the models are evaluated on the corresponding balanced testing or validation datasets. Top-1 accuracy is adopted for all the comparisons, denoted as *All*. In order to better analyze the performance on classes with different sampling frequencies, we further report the accuracy on three-class subsets: *Many-shot* (more than 100 samples), *Medium-shot* ($20 \sim 100$ samples) and *Few-shot* (less than 20 samples), following the work in [21].

4.3. Experimental Results

To verify the effectiveness of our approach, we mainly compared to the following methods. **CE:** training with a cross-entropy loss [16]; **Balanced Softmax:** BS [33], Logit adjust [29], DisAlign [17], LADE [51], ALA Loss [53]; **Contrastive Learning:** PaCo [8], KCL [20], Hybrid-SC [36]; **Optimal Transport:** OTLM [31]; **Other Methods:** Focal Loss [26], cRT [21], BBN [55], Bag of Tricks [52], MARC [45].

Experimental results on long-tailed CIFAR. We conduct extensive experiments on CIFAR-10/100 with imbalance factors of 200, 100, and 10 with the same set of the work of [33]. We mainly compare to the current SOTA method BALMS [33]. The experimental results are summarized in Table 1. As a baseline algorithm, a network is trained by minimizing the empirical cross-entropy loss without regularization. Compared with the previous methods, the performance of our method is greatly improved. Specifically, our method outperforms the SOTA methods by 2.6%, 1.8% and 1.8% on CIFAR100-LT with imbalance factors 200, 100, and 10, respectively. Our method also surpasses the SOTA methods by 1.9%, 1.0%, and 1.4% on CIFAR10-LT under imbalance factors 200,100 and 10, respectively, which testify the effectiveness of our method.

Experimental results on large-scale datasets. Here, we mainly compare our algorithm with the SOTA method ALA Loss and the correlational method Balance softmax.

ImageNet-LT The experimental results with ResNet-50 and ResNeXt-50 on ImageNet-LT are reported in Table 2.

As shown in Table 2, our method achieves superior performance to existing methods on both networks. Comparing our method with state-of-the-art ALA loss, the Top-1 accuracy has been improved by 0.9% and 1.3% with ResNet-50 and ResNeXt-50, respectively. Compared with other Balance softmax losses, our method gets better results on all three subsets, showing the comprehensive advantages of our method. Unlike other methods that improve tail classes at the sacrifice of head classes, our method not only achieves better results on tail classes, but also achieves comparable results on head classes to cross cntropy loss.

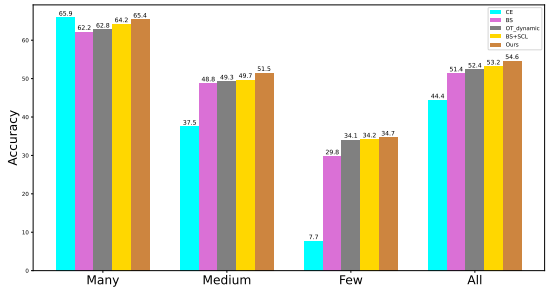

Figure 3. Histogram of accuracy for ablation studies of OT-dynamic softmax loss on the ImageNet-LT with ResNeXt-50. All models are trained in 90 epochs.

iNaturalist 2018 and Places-LT. The experimental results on iNaturalist 2018 and Places-LT are reported in Table 3. On the real-world long-tailed iNaturalist 2018, our method again outperforms other methods, especially on the medium-shot subset. Compared with the SOTA method ALA Loss, our method achieves the best overall accuracy, with more than 0.4% performance gain and 1.2% gain on the medium-shot subset. For Places-LT, we observe a similar improvement. Our method achieves the best overall accuracy of 40.5%, which is 0.4% higher than ALA Loss.

4.4. Ablation study

Quantitative analysis. In this section, we conduct a series of experiments to examine the effect of each component in our method. Fig. 3 shows the histogram of accuracy on four class subsets with five compared methods. According to the comparisons shown in Fig. 3, we have the following observations:

- OT-dynamic Softmax aims to dynamicly adjust the decision boundary in the training phase, which can boost the performance of the long-tailed classification to a certain extent. Compared with BS (orchid bar), OT-dynamic (gray bar) achieves considerable better results on all subsets. Moreover, it especially improves the accuracy in the few-shot subset by 4.3%, indicating the advantage of tackling the long-tailed problem from the semi-optimal transport perspective.
- SCL purposes to distribute the features evenly in the feature space so that more samples fall in the correct cell decomposition. The comparison between BS and BS + SCL (gold bar) reveals that SCL is able to improve overall performances. What's more, SCL brings significant gains on the many-shot subset and few-shot subset.
- According to the peru bar, the combination setting OT-dynamic + SCL (Ours) achieves the best result. Compared

Method	Pub.	CIFAR-10-LT			CIFAR-100-LT		
		200	100	10	200	100	10
CE	CVPR′16	71.2	77.4	90.0	41.0	45.3	61.9
Focal Loss	ICCV′17	71.8	77.1	90.3	40.2	43.8	60.0
BBN	CVPR′20	-	79.8	88.3	-	42.5	59.1
cRT	ICLR′20	76.6	82.0	91.0	44.5	50.0	63.3
BS	NeurIPS′20	81.5	84.9	91.3	45.5	50.8	63.0
LADE	CVPR′21	-	-	-	-	45.4	61.7
Hybrid-SC	CVPR′21	-	81.4	91.1	-	46.7	63.0
Bag of Tricks	AAAI′21	-	80.0	-	-	47.8	-
PaCo	ICCV′21	-	-	-	-	52.0	64.2
MARC	ArXiv′21	81.1	85.3	-	47.4	50.8	-
OTLM	ICLR′22	-	-	-	37.8	42.6	61.2
Our method	-	**83.4**	**86.3**	**92.7**	**50.0**	**53.8**	**66.4**

Table 1. Top-1 accuracy (%) comparison on CIFAR-10/100-LT with ResNet-32 for different imbalance factors.

Method	Pub.	ResNeXt-50				ResNet-50			
		Many	Medium	Few	All	Many	Medium	Few	All
CE	CVPR′16	**65.9**	37.5	7.7	44.4	**64.0**	33.8	5.8	41.6
Focal Loss	ICCV′17	64.3	37.1	8.2	43.7	-	-	-	-
BBN	CVPR′20	-	-	-	49.3	-	-	-	48.3
cRT	ICLR′20	61.8	46.2	27.4	49.6	58.8	44.0	26.1	47.3
BS	NeurIPS′20	62.2	48.8	29.8	51.4	61.0	47.0	27.2	49.7
Logit adjust	ICLR′20	-	-	-	-	-	-	-	51.1
DisAlign	CVPR′21	61.5	50.7	33.1	52.6	59.9	49.9	31.8	51.3
LADE	CVPR′21	62.3	49.3	31.2	51.9	59.9	49.9	31.8	51.3
KCL	ICLR′21	-	-	-	-	61.8	49.4	30.9	51.5
PaCo	ICCV′21	59.7	**51.7**	36.5	52.7	-	-	-	-
MARC	ArXiv′21	60.4	50.3	**36.6**	52.3	-	-	-	-
ALA Loss	AAAI′22	64.1	49.9	34.7	53.3	62.4	49.1	**35.7**	52.4
OTLM	ICLR′22	-	-	-	-	-	-	-	52.4
Our method	-	65.4	51.5	34.7	**54.6**	63.8	**50.6**	33.5	**53.3**

Table 2. Performance comparison with state-of-the-art methods on ImageNet-LT with ResNeXt-50 and ResNet-50.

with CE (blue bar), Ours obtain pretty better results on both medium-and few-shot subsets, with only a slight decline on many-shot subset.

In addition, Table 4 also shows that our method gets the best results on CIFAR-10/100-LT. It is consistent with our design principle. That is, OT-dynamic Softmax pay more attention to adjust the decision boundary, SCL focus more on features uniformity.

Qualitative analysis. In this section, we conduct qualitative analysis to characterize our OT-dynamic softmax intuitively and comprehensively. Specifically, we further visualize and analyze the advantages of OT-dynamic softmax from the perspectives of avoiding overfitting in a few-shot classes and adjusting decision boundaries.

Method	Pub.	Places-LT				iNaturalist2018			
		Many	Medium	Few	All	Many	Medium	Few	All
CE	CVPR′16	**45.7**	27.4	8.2	30.2	**72.7**	63.8	58.7	61.7
Focal Loss	ICCV′17	41.1	34.8	22.4	34.6	-	-	-	-
BBN	CVPR′20	-	-	-	-	49.4	70.8	65.3	66.3
cRT	ICLR′20	42.0	37.6	24.9	36.7	69.0	66.0	63.2	65.2
BS	NeurIPS′20	41.2	39.8	31.6	38.7	-	-	-	69.8
Logit adjust	ICLR′20	-	-	-	-	-	-	-	66.4
DisAlign	CVPR′21	40.4	**42.4**	31.8	39.3	-	-	-	70.6
LADE	CVPR′21	42.8	39.0	31.2	38.8	-	-	-	70.0
KCL	ICLR′21	-	-	-	-	-	-	-	68.6
MARC	ArXiv′21	39.9	39.8	32.6	38.4	-	-	-	70.4
ALA Loss	AAAI′22	43.9	40.1	32.9	40.1	71.3	70.8	**70.4**	70.7
Our method	-	42.7	40.4	**36.7**	**40.5**	70.0	**72.0**	70.2	**71.1**

Table 3. Performance comparison with state-of-the-art methods on Places-LT and INaturalist2018 with ResNet-152 and ResNet-50.

Method	**BS**	**SCL**	$\nabla \mathbf{E}(\boldsymbol{h})$	CIFAR-10-LT			CIFAR-100-LT		
				200	100	10	200	100	10
CE	✗	✗	✗	71.2	77.4	90.0	41.0	45.3	61.9
BS	✔	✗	✗	81.5	84.9	91.3	45.5	50.8	63.0
SCL	✗	✔	✗	76.0	82.4	91.6	43.5	47.4	63.7
BS+SCL	✔	✔	✗	82.3	85.6	92.0	48.2	52.9	65.5
OT-dynamic	✔	✗	✔	82.6	84.5	91.4	46.1	51.7	64.1
Our method	✔	✔	✔	**83.4**	**86.3**	**92.7**	**50.0**	**53.8**	**66.4**

Table 4. **Ablation studies for OT-dynamic Softmax Loss.** Results on the test set of CIFAR10/100-LT with ResNet-32. The first line is the results of Cross Entropy (CE); the last line is our method. **OT-dynamic Softmax** is denoted in Equation (15).

To intuitively reflect the advantage of our OT-dynamic softmax over fixd class frequency based Balanced Softmax methods, we visualize the accuracy curve during training. According to the performance comparisons shown in Fig. 4, we obetain the following observations:

• The "All" and "Many" frames show that the curve train-OT and test-OT are always above curve trian-BS and test-BS, respectively, which indicates OT-dynamic softmax can significantly improve the predicted accracies compared with BS.

• In the "Medium" and "Few" frames, the curve train-BS is over curve train-OT, while the curve test-BS is below curve test-OT. This means overfitting in the BS training, especially in the few-shot subset. In contrast, our OT-dynamic softmax is able to dynamically adjust the decision boundaries to avoid this problem.

• The fluctuation of curve test-OT is significantly smaller than curve test-BS, which suggests that our OT-dynamic softmax enables a more stable training process compared to BS.

To gain additional insight, we look at the t-SNE projection of learned features and compared CE, BS, and SCL with our proposed OT-dynamic softmax loss. Fig. 5 shows that the cell decomposition in our learned feature space is more uniform and the decision boundaries are clearer, which is particularly evident in the red class.

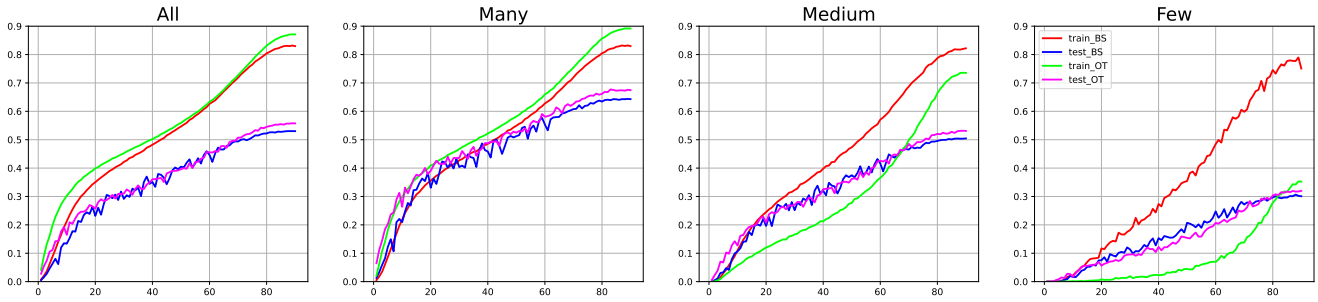

Figure 4. Performance comparison of BS and OT-dynamic softmax in the training process. We use ResNext-50 for end-to-end training on ImageNet-LT with 90 epochs. The x-axis is the number of training epoch and the y-axis is the classification accuracy.

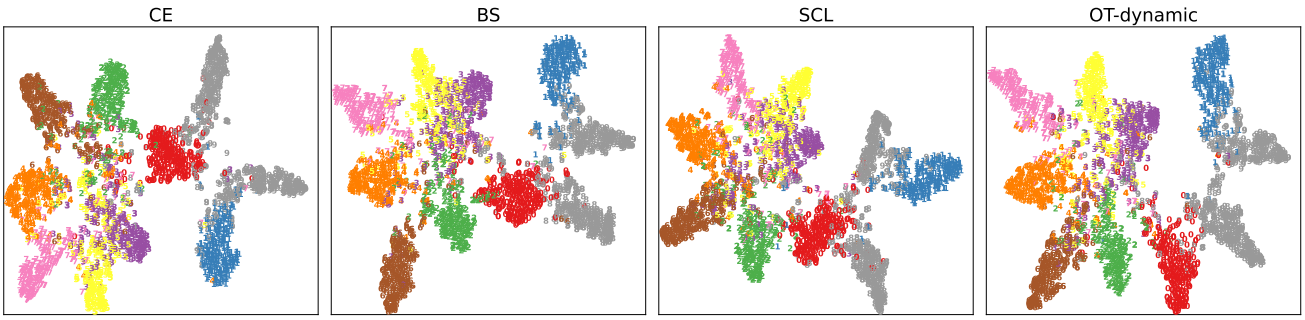

Figure 5. t-SNE visualization of feature space of CIFAR-10-LT with imbalance factor 100 obtained using different methods.

5. Conclusions

In this paper, we analyze the long-tailed classification problem from the perspective of semi-discrete optimal transmission, i.e., the feature space is considered a continuous source domain, and the classifier weights are considered discrete target points. Our analysis shows that the imbalance in the data leads to the decision boundary being biased toward classes with low frequency. Therefore, we propose an optimal transport dynamic softmax loss to adjust the decision boundary dynamically. Furthermore, our method combined the supervised contrastive loss to allow the feature space to satisfy the uniform distribution. Extensive and comprehensive experimental results show that our method outperforms the existing SOTA methods on widely used long-tailed benchmarks, including CIFAR-10/100-LT, ImageNet-LT, iNaturalist 2018, and Places-LT. Moreover, we note that OT-dynamic softmax loss can avoid overfitting in the tail classes and improve the robustness of the training process.

Acknowledgement

This work was partially supported by National Key R&D Program of China 2021YFA1003003 and National Natural Science Foundation of China under Grant No. 61936002, T2225012.

References

[1] D. An, Y. Guo, N. Lei, Z. Luo, S.-T. Yau, and X. Gu. Ae-ot: a new generative model based on extended semi-discrete optimal transport. *ICLR 2020*, 2019. 1, 3

[2] Y. Brenier. Polar decomposition and increasing rearrangement of vector-fields. *COMPTES RENDUS DE L ACADEMIE DES SCIENCES SERIE I-MATHEMATIQUE*, 305(19):805–808, 1987. 3

[3] M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 1, 2

[4] J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019. 2

[5] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *arXiv preprint arXiv:1906.07413*, 2019. 6

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6

[7] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in Neural Information Processing Systems*, 30, 2017. 3

[8] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF interna-*

tional conference on computer vision, pages 715–724, 2021. 1, 2, 7

[9] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4109–4118, 2018. 1

[10] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019. 3

[11] Z. Ge, S. Liu, Z. Li, O. Yoshie, and J. Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 1

[12] X. Gu, F. Luo, J. Sun, and S.-T. Yau. Variational principles for minkowski type problems, discrete optimal transport, and discrete monge-ampere equations. *arXiv preprint arXiv:1302.5472*, 2013. 2, 3

[13] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 3

[14] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 2

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 7

[17] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 1, 2, 6, 7

[18] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002. 1, 2

[19] K. Kandasamy, W. Neiswanger, J. Schneider, B. Poczos, and E. P. Xing. Neural architecture search with bayesian optimisation and optimal transport. *Advances in neural information processing systems*, 31, 2018. 3

[20] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 1, 2, 7

[21] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1, 2, 6, 7

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 2, 6

[23] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009. 2, 6

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[25] N. Lei, K. Su, L. Cui, S.-T. Yau, and X. D. Gu. A geometric view of optimal transportation and generative model. *Computer Aided Geometric Design*, 68:1–21, 2019. 1, 3

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 7

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[28] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 6

[29] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2020. 1, 2, 7

[30] L. Mi, W. Zhang, X. Gu, and Y. Wang. Variational wasserstein clustering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 322–337, 2018. 1

[31] H. Peng, M. Sun, and P. Li. Optimal transport for long-tailed recognition with learnable cost matrix. In *International Conference on Learning Representations*, 2021. 1, 3, 7

[32] S. T. Rachev and L. Rüschendorf. *Mass transportation problems: Volume I: theory*, volume 1. Springer Science & Business Media, 1998. 3

[33] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, and H. Li. Balanced meta-softmax for long-tailed visual recognition. *arXiv preprint arXiv:2007.10740*, 2020. 1, 2, 6, 7

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1, 2, 6

[35] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018. 3

[36] D. Samuel and G. Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9495–9504, 2021. 1, 2, 7

[37] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 1, 2

[38] Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015. 3

[39] K. S. Tai, P. Bailis, and G. Valiant. Sinkhorn label allocation: Semi-supervised classification via annealed self-training. *arXiv preprint arXiv:2102.08622*, 2021. 1

[40] K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020. 1

[41] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 2, 6

[42] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021. 1, 2

[43] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 2

[44] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020. 2

[45] Y. Wang, B. Zhang, W. Hou, Z. Wu, J. Wang, and T. Shinozaki. Margin calibration for long-tailed visual recognition. *arXiv preprint arXiv:2112.07225*, 2021. 7

[46] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7032–7042, 2017. 1

[47] X.-S. Wei, P. Wang, L. Liu, C. Shen, and J. Wu. Piecewise classifier mappings: Learning fine-grained learners for novel categories with few examples. *IEEE Transactions on Image Processing*, 28(12):6116–6125, 2019. 6

[48] L. Xiang, G. Ding, and J. Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020. 2

[49] H. Xu, D. Luo, and L. Carin. Scalable gromov-wasserstein learning for graph partitioning and matching. *Advances in neural information processing systems*, 32, 2019. 3

[50] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu. Semi-supervised optimal transport for heterogeneous domain adaptation. In *IJCAI*, volume 7, pages 2969–2975, 2018. 3

[51] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 1, 2, 7

[52] Y. Zhang, X.-S. Wei, B. Zhou, and J. Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3447–3455, 2021. 7

[53] Y. Zhao, W. Chen, X. Tan, K. Huang, and J. Zhu. Adaptive logit adjustment loss for long-tailed visual recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3472–3480, 2022. 7

[54] Q. Zhong, C. Li, Y. Zhang, H. Sun, S. Yang, D. Xie, and S. Pu. Towards good practices for recognition & detection. In *CVPR workshops*, volume 1, 2016. 2

[55] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 2, 6, 7

[56] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 2, 6