Robust Visual Recognition with Class-Imbalanced Open-World Noisy Data

Na Zhao[1*], **Gim Hee Lee**[2]

[1]Singapore University of Technology and Design
[2]National University of Singapore

Abstract

Learning from open-world noisy data, where both closed-set and open-set noise co-exist in the dataset, is a realistic but underexplored setting. Only recently, several efforts have been initialized to tackle this problem. However, these works assume the classes are balanced when dealing with open-world noisy data. This assumption often violates the nature of real-world large-scale datasets, where the label distributions are generally long-tailed, *i.e.* class-imbalanced. In this paper, we study the problem of robust visual recognition with class-imbalanced open-world noisy data. We propose a probabilistic graphical model-based approach: *i*MRF to achieve label noise correction that is robust to class imbalance via an efficient iterative inference of a Markov Random Field (MRF) in each training mini-batch. Furthermore, we design an agreement-based thresholding strategy to adaptively collect clean samples from all classes that includes corrected closed-set noisy samples while rejecting open-set noisy samples. We also introduce a noise-aware balanced cross-entropy loss to explicitly eliminate the bias caused by class-imbalanced data. Extensive experiments on several benchmark datasets including synthetic and real-world noisy datasets demonstrate the superior performance robustness of our method over existing methods. Our code is available at https://github.com/Na-Z/LIOND.

1 Introduction

The success of modern deep neural networks for visual recognition heavily rely on the availability of large-scale well-annotated datasets such as ImageNet. Unfortunately, the extreme cost and difficulty in annotating extensive data severely limits the construction of more large-scale datasets with precise annotations. This limitation impedes the development of deep models for fine-grained visual recognition. In contrast, tremendous data with noisy labels can be easily accessed from the online search engines. Hence, designing robust visual recognition approaches that are unbiased to the noise data is appealing. This is commonly known as *learning with noisy labels* in the literature (Song et al. 2022).

Many studies on learning from noisy data (Han et al. 2018; Jiang et al. 2018; Li, Socher, and Hoi 2020; Li et al. 2019; Liu et al. 2020b; Patrini et al. 2017; Wei et al. 2020; Zhang and Sabuncu 2018) generally follow the closed-set assumption, *i.e.* the real labels of the noisy samples (samples with incorrect labels) are within the same label space as the clean samples. An example of **closed-set** (*a.k.a.* in-distribution) **noisy data** is shown in Fig. 1(a), where one image of 'cat' is mislabeled as 'dog' while two images of 'dog' are mislabeled as 'cat' and 'horse', respectively. Although these prior works have achieved impressive progress, their closed-set assumption is unrealistic in many real applications. This is because the training label space is usually a subset of the labels in real-world, and thus it is inevitable to take samples that are not belonging to the training label space when we collect data from the open world. This kind of samples is usually known as open-set (*a.k.a.* out-of-distribution) noisy data (Wang et al. 2018; Wei et al. 2021). The prevalent co-existence of closed-set and open-set noise in the training data, which is also referred as the **open-world noisy data** as illustrated in Fig. 1(b), is a practical problem. However, it has not been studied until very recently (Li, Xiong, and Hoi 2021; Sachdeva et al. 2021; Wu et al. 2021; Yao et al. 2021).

Despite the relatively realistic open-world setting compared to the naive closed-set assumption, existing approaches under this setting still suffer from an unrealistic class balance assumption that hypothesizes a uniform density of the label distribution. This balance assumption violates the real-world large-scale datasets (Li et al. 2017; Van Horn et al. 2018) that generally follow long-tailed distributions (Reed 2001) with severe class imbalance between the head and tail classes. To carry out robust visual recognition using real-world large-scale datasets, it is imperative to train a robust deep neural network that can effectively learn from the **class-imbalanced open-world noisy data**, as illustrated in Fig. 1 (c).

This paper is the first to explicitly study this unexplored yet pragmatic direction. The challenges lie in the difficulty of correctly identifying clean samples, closed-set noisy samples, and scarce samples from the tail classes. Prior works that use either temporally averaged model (Yao et al. 2021) or smooth neighbors (Li, Xiong, and Hoi 2021) for open-world noisy label correction face challenges with class imbalance. JoSRC (Yao et al. 2021) may lead to overfitting on noisy labels, favoring head classes, while ProtoMix(Li, Xiong, and Hoi 2021) may struggle to form meaningful neighborhoods due to imbalanced class distributions. In contrast, NGC (Wu et al. 2021) that constructs a global graph with all training samples

*Corresponding author: na_zhao@sutd.edu.sg.

Figure 1: Exemplar comparison of different problem settings for noisy data learning. This toy dataset only contains three categories: cat, dog, and horse. Tags on the images indicate the given labels. Green, blue, and orange highlights the clean sample, closed-set noisy sample, and open-set noisy sample, respectively.

and employs label propagation (Iscen et al. 2019; Zhou et al. 2003) over the graph is less susceptible to the class imbalance issue. However, the label propagation process over the entire graph becomes increasingly memory- and time-consuming when the size of graph increases, which impedes its feasible use in practice. Moreover, all these prior works reply on two fixed thresholds to select clean samples or reject open-set noisy samples. Such a fixed thresholding strategy is sub-optimal since it cannot be dynamically adjusted based on the learning status. This can lead to the inaccurate acceptance of clean samples and wrong rejection of open-set noisy samples.

We propose a probabilistic graphical model-based approach, which we called iterative Markov Random Field (iMRF) to achieve label noise correction for class-imbalanced data. Our iMRF represents the true labels of all training samples as a set of latent random variables and performs inference in each training mini-batch efficiently. The MRF is formed by the K-nearest neighbors (KNN) graph of the embeddings from all training samples at each epoch. At each iteration, iMRF infers each latent random variable in the batch based on its predicted value from the current model and the inferred values of its KNN from the previous iteration. The designed smoothness energy function (*c.f.* Sec. 3.1 for the details) in our iMRF is able to take care of the samples in the tail classes by weighing down the contributions of their dissimilar neighbors. Furthermore, we design an agreement-based thresholding strategy that can adaptively adjust the threshold for selecting clean samples from all in-distribution classes or rejecting open-set noisy samples. Specifically, we introduce a tripartite agreement mechanism to obtain clean samples with high confidence at each epoch. The tripartite agreement is met by the consensus among the given label, current inference from iMRF, and previous inference from iMRF. Once we have the high-confident clean samples *w.r.t.* current epoch, two adaptive thresholds are set accordingly to collect clean samples used for next epoch. Inspired by balanced softmax (Ren et al. 2020), we also present a noise-aware balanced cross-entropy loss to explicitly eliminate the bias of class-imbalanced data. By training the model with this loss and a set of losses that includes two types of contrastive losses, we expect the model to be robust to the class-imbalanced open-world noisy data. The **main contributions** of this work are summarized as follows: 1) We propose a probabilistic graphical model-based method iMRF to effectively correct label noise with semi-global smoothness and efficient inference. 2) We design an agreement-based thresholding strategy that can adaptively adjust the thresholds for sample selection. Our proposed tripartite agreement mechanism is aware of both the model learning status and the given and previously inferred labels. 3) We conduct extensive experiments on both synthetic and real-world noisy datasets with a variety of experimental settings. Our proposed method consistently shows superior performance over existing state-of-the-art methods.

2 Related Work

Learn from Noisy Data. A huge body of approaches (Zhang et al. 2018; Xia et al. 2020; Yao et al. 2020; Zhang, Xing, and Liu 2021; Zhang et al. 2021b; Huang et al. 2019) have been contributed to the filed of learning from noisy data, which can be roughly summarized as: a) *robust regularization* including network regularization (*e.g.*, dropout, weight decay, and batch normalization) and data augmentation such as mixup, b) *loss adjustment* including loss correction and loss reweighting, c) *label correction*, and d) *sample selection*. Notably, recent start-of-the-art approaches such as DivideMix (Liu et al. 2020b) and ELR+ (Li, Socher, and Hoi 2020) are mainly hybrid approaches that combines different techniques, *e.g.* sample selection and semi-sueprvised learning.

Despite the progress of these approaches in learning from noisy data, their closed-set assumption largely limits their applications in large-scale real-world datasets where open-set and closed-set noise co-exist. The problem of learning from open-set noisy data is firstly unveiled by ILON (Wang et al. 2018), it iteratively selects the noisy labels via the density estimation within neighorhood and learns discriminative features by enlarging the distance between clean and noisy samples

with a contrastive loss. Since then, this practical problem has attracted more attentions (Li, Xiong, and Hoi 2021; Sachdeva et al. 2021; Wu et al. 2021; Yao et al. 2021). Most of them adopt the iterative learning strategy similar as ILON. They gradually select more clean samples via label correction and subsequently perform semi-supervised learning with the selected clean samples as labeled and the remaining ones as unlabeled. All methods use contrastive learning for unlabeled samples, but differ in label correction and sample selection approaches. Jo-SRC (Yao et al. 2021) uses JS-divergence to measure clean sample likelihood and prediction disagreement between augmented views for open-set sample likelihood. It employs two thresholds for clean and open-set sample selection. Then, it uses the temporally averaged model to correct labels for closed-set noisy samples. ProtoMix (Li, Xiong, and Hoi 2021) cleans the closed-set noise by aggregating predictions of neighboring samples. Unlike the local smoothness manner in ProtoMix, NGC (Wu et al. 2021) adopts the label propagation technique that can take account of both local and global smoothness. However, all of these approaches follow the class-balanced assumption, which violates the long-tailed nature of large-scale datasets. Moreover, their techniques face either failures or heavy computation issue when dealing with class-biased large-scale real-world datasets.

Learning with Imbalanced Classes. A large and growing body of literature (Zhang et al. 2021a) has investigated on learning with imbalanced classes, *a.k.a.* long-tailed learning. The goal is to learn a unbiased deep neural network given a class-biased training dataset, where some classes have massive samples (*i.e.* head classes) yet some are with very few samples (*i.e.* tail classes). Recent techniques (Zhou et al. 2020; Ren et al. 2020; Liu et al. 2020a; Xiang, Ding, and Han 2020) tackle this problem mainly from three perspectives: 1) *data re-sampling* that manipulates training samples via over-sampling, under-sampling, or class-balanced sampling to produce a balanced distribution, 2) *loss adjustment* that employs class-level re-weighting to regulate the influence of label distribution on loss weights or encourage larger margins between features and classifier for tail classes, and 3) *transfer learning* that allows for various transferring schemes, such as head-to-tail knowledge transfer and knowledge distillation. Among these technique, the class-level re-weighting is a simple but effective method. It can be applied to losses or model predictions, resulting in weighted softmax loss (Kang et al. 2019) and balanced softmax (Ren et al. 2020), respectively.

Unfortunately, these existing approaches are not robust to real applications since they assume all the data are clean, which is hard to be hold in the real data. Thus, the performance of these approaches, especially on tail classes, can greatly deteriorate when the class imbalance issue occurs with open-world noisy data. Our paper targets on this practical yet challenging problem, and presents a noise-aware balanced cross-entropy loss to explicitly eliminate the bias of class-imbalanced data.

3 Our Method

In robust visual recognition with class-imbalanced open-world noisy data, we are given a training set $\mathfrak{D} = \{\mathbf{x}_i, \hat{y}_i\}_{i=1}^N$, where $\mathbf{x}_i$ is the i-th training image and $\hat{y}_i \in \{1, ..., C\}$ is its given label. Let n^k be the number of samples of the k-th class, we have $\sum_{k=1}^C n^k = N$ and $n^i \neq n^j$. Since the training data is long-tailed, we have the i-th class as a head class and the j-th class as a tail class with $n^i \gg n^j$. Let $y_i^\star$ denote the real label of the sample $\mathbf{x}_i$, the label of a sample is clean when $\hat{y}_i = y_i^\star$, and noisy when $\hat{y}_i \neq y_i^\star$. Specifically, the label is a closed-set noise if $y_i^\star \in \{1, ..., C\}$, otherwise it is a open-set noise. Our objective is to learn a robust model f_Θ, which can recognize the unseen testing data with high accuracy from the class-imbalanced open-world noisy training data.

We adopt the popular iterative learning strategy in previous works (*e.g.* ProtoMix and NGC) that consists of two iterative steps: 1) **Label correction-based sample selection.** We gradually correct the noisy labels via our proposed iMRF (Sec. 3.1) and then select the clean and corrected samples (Sec. 3.2), *i.e.* to reject the labels of the open-set samples. 2) **Robust representation learning.** We train the model to learn robust representation using both supervised and unsupervised losses (Sec. 3.3) with the selected clean and corrected samples as labeled. The two steps are carried out alternatively until the model converges.

3.1 Label Correction with iMRF

We introduce iMRF, a probabilistic graphical model-base method, for label noise correction. In iMRF, real labels and model predictions are represented as latent and observed random variables in a Markov random field (MRF). A batch-update inference algorithm is derived to infer the real labels of all samples. Intuitively, neighboring samples, close in the well-learned feature space, are expected to share similar labels. This is commonly known as the *smoothness constraint* in the literature of markov random field (Koller and Friedman 2009), semi-supervised learning (Iscen et al. 2019; Zhou et al. 2003), *etc*. Meanwhile, model predictions, anticipated to provide reliable information on the real label (Reed et al. 2014) as the model becomes more robust and resilient to noise and class imbalance during training, are considered as observed random variables. The constraint enforcing that the latent random variable remains close to these predictions is referred to as the *fitness constraint* in our iMRF. By integrating the smoothness and fitness constraints, our iMRF aims to estimate the posterior probability of all real labels $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, given all data samples $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and model predictions $O = \{\mathbf{o}_1, ..., \mathbf{o}_N\}$:

$$P(Y \mid X, O, \Theta) = \frac{1}{Z} \exp\Big\{ -\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \mathbf{A}_{ij} \cdot E_s(\mathbf{y}_i, \mathbf{y}_j) - \sum_{i=1}^{N} E_f(\mathbf{y}_i, \mathbf{o}_i) \Big\}, \tag{1}$$

where $\mathbf{A} \in \{0, 1\}^{N \times N}$ is a sparse adjacency matrix that represents a KNN graph formed by connecting each sample to its K nearest neighbors in the learned feature space. $\mathbf{y}_i$ is a latent random variable, denoting the real label as a one-hot vector in $\{0, 1\}^C$. $\mathbf{o}_i \in \mathbb{R}^C$ is an observed random variable, denoting the categorical distribution predicted by the model f_Θ. Θ represents the learnable parameters of the model f_Θ.

E_s and E_f denote the smoothness and the fitness energy function, respectively. Z is the partition function.

For each pair of samples $\mathbf{x}_i$ and $\mathbf{x}_j$, the smoothness energy function $E_s(\mathbf{y}_i, \mathbf{y}_j)$ measures the disagreement between $\mathbf{y}_i$ and $\mathbf{y}_j$, weighted by an adaptive cost coefficient that is aware of the distance between the learned feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$. The formulation of E_s is as follow:

$$E_s(\mathbf{y}_i, \mathbf{y}_j) = \mathbb{1}[\mathbf{x}_j \in \mathfrak{D}_s \cup \mathfrak{D}_w] \cdot J(\mathbf{f}_i, \mathbf{f}_j) \cdot \|\mathbf{y}_i - \mathbf{y}_j\|_2, \quad (2)$$

and

$$J(\mathbf{f}_i, \mathbf{f}_j) = \alpha \cdot \exp(\beta \cdot D(\mathbf{f}_i, \mathbf{f}_j)), \quad (3)$$

where the indicator function $\mathbb{1}[\cdot] = 0$ if the j-th neighbor is not selected as the clean sample (*c.f.* Eq. 5 for the definitions of $\mathfrak{D}_s$ and $\mathfrak{D}_w$). By adding this constraint, those potential outliers (*i.e.*, open-set noisy samples) can be ruled out from the computation of the smoothness energy. Such a strategy would lead to a more accurate inference by alleviating the adverse impact of open-set noise especially on tail classes that are hard to be distinguished from the open-set noise. $J(\cdot, \cdot)$ is a non-negative increasing function given by the cosine distance $D(\mathbf{f}_i, \mathbf{f}_j)$. α and β are hyper-parameters. The intuition behind J is that the contribution of the disagreement between two neighboring samples should be proportional to their similarity in the feature space. For example, let us consider two neighbors $\mathbf{y}_j$ and $\mathbf{y}_k$ of the random variable $\mathbf{y}_i$. The case where $\mathbf{x}_k$ is closer to $\mathbf{x}_i$ than $\mathbf{x}_j$ in the feature space, *i.e.* $D(\mathbf{f}_i, \mathbf{f}_k) \leq D(\mathbf{f}_i, \mathbf{f}_j)$ should result in $E_s(\mathbf{y}_i, \mathbf{y}_k) \leq E_s(\mathbf{y}_i, \mathbf{y}_j)$.

For each sample $\mathbf{x}_i$, the fitness-based energy function $E_f(\mathbf{y}_i, \mathbf{o}_i)$ aims to measure the divergence between the state of $\mathbf{y}_i$ and the model prediction $\mathbf{o}_i$, which is formulated as:

$$E_f(\mathbf{y}_i, \mathbf{o}_i) = \gamma \cdot \|\mathbf{y}_i - \mathbf{o}_i\|_2, \quad (4)$$

where γ is a hyper-parameter that represents the cost coefficient for the fitness energy.

Unfortunately, the exact inference of the posterior distribution $P(Y|X, O, \Theta)$ in Eq. 1 is intractable due to the high dimensionality of the latent random variable Y and high computation cost of Z in Eq. 1. One naive solution is to assume complete independence of the latent random variables and ignore the pairwise smoothness terms $E_s(\mathbf{y}_i, \mathbf{y}_j)$. However, this leads to a trivial solution where label smoothness is ignored. We can also do approximate inference with variational inference (Jordan et al. 1999), MCMC Sampling (Andrieu et al. 2003), alpha-expansion (Boykov, Veksler, and Zabih 2001), *etc*. However, these approximate inference methods are usually too slow to be computed during training. In view of these limitations, we derive a batch-update inference algorithm with a computation complexity and accuracy between the naive and approximate inference algorithms.

Our batch-update algorithm evaluates the posterior probabilities of a subset of samples (*i.e.* batch) at a time. For each latent random variable during an evaluation step, we fix its K nearest neighbors at their inferred values from the previous evaluation. Note that the complexity of our batch-update algorithm is dependent only on the batch size and number of nearest neighbors K. The initial values of latent random variables Y are set as the model predictions after a warm-up training stage, except for those chosen as clean samples by our thresholding strategy. We set the values of these selected clean samples to the one-hot vector of the given labels.

We then use the computed posterior probability $P(Y \mid X, O, \Theta)$ for the sample selection in the next section. Furthermore, the resultant $\tilde{y}_i = \arg\max \mathbf{p}_i$ are used as the corrected label for $\mathbf{x}_i$, where $\mathbf{p}_i$ represents $P(\mathbf{y}_i \mid X, O, \Theta)$.

3.2 Tripartite Agreement for Sample Selection

Similar to prior works (Li, Xiong, and Hoi 2021; Wu et al. 2021; Yao et al. 2021) in learning open-world noisy data, we construct a mixture of *strongly labeled* set $\mathfrak{D}_s$, *weakly labeled* set $\mathfrak{D}_w$, and *unlabeled* set $\mathfrak{D}_u$ for training the model. The samples in the strongly labeled set contains their given labels, while the samples in the weakly labeled set are assigned the corrected labels. Generally, we expect the strongly labeled set to include clean samples and the unlabeled set to include the open-set noisy samples. These prior works utilize two manually defined and fixed thresholds (*i.e.* η_l and η_h) to identify the three sets. Concretely, the thresholds η_l and η_h are used for accepting the clean samples and the corrected closed-set noisy samples, respectively. Since the model performance may change along the training, such a fixed thresholding strategy that cannot be adjusted with the learning status of the model is not optimal. To this end, we design an agreement-based thresholding strategy to take the model learning status into account. Specifically, we propose a tripartite agreement mechanism to check the consensus among the given labels $\{\hat{y}_i\}$, current corrected labels $\{\tilde{y}_i^t\}$, and previous corrected labels $\{\tilde{y}_i^{t-1}\}$. We postulate that those samples that fulfil the tripartite agreement are clean samples with high confidence, and thus can be used as a *reference set* for determining the thresholds, *i.e.* $\mathcal{I}_{ref} \leftarrow i$, if $\hat{y}_i = \tilde{y}_i^t = \tilde{y}_i^{t-1}\ \forall i = 1, \cdots, N$. Since the current and previous corrected labels are derived from the same model at different training epochs, they inherently reflect the learning status of the model.

Consequently, we get the posterior probabilities of the samples in the reference set at their given labels and then rank them in an ascending order, *i.e.* $\mathcal{S}_{\text{sort}} \leftarrow \text{AscendSort}(\{\mathbf{p}_i(y_i) : y_i \in \mathfrak{D},\ i \in \mathcal{I}_{ref}\})$. In view of the high confidence of the reference set, we believe the obtained probabilities can manifest the true probabilities distribution *w.r.t.* the real labels. A straightforward way is to select the minimum and maximum probabilities as the thresholds η_l^* and η_h^*, respectively. Nonetheless, the reference set is impossible to be completely clean. To be tolerant to the potential noise in the reference set, we select the top $\epsilon_l\%$ and the last $\epsilon_h\%$ probabilities in $\mathcal{S}_{\text{sort}}$ and use the corresponding mean as η_l^* and η_h^*, respectively. ϵ_l and ϵ_h are hyper-parameters to control the percentage of samples that are probably outliers. Finally, the resultant training data including aforementioned three sets can be collected as:

$$\mathfrak{D}^\diamond = \underbrace{\{\mathbf{x}_i, \hat{y}_i \mid \mathbf{p}_i(\hat{y}_i) > \eta_l^*\}}_{\text{strongly labeled}} \cup \underbrace{\{\mathbf{x}_i, \tilde{y}_i \mid \max(\mathbf{p}_i) > \eta_h^*\}}_{\text{weakly labeled}} \cup \underbrace{\{\mathbf{x}_i \mid \mathbf{p}_i(\hat{y}_i) \leq \eta_l^* \vee \max(\mathbf{p}_i) \leq \eta_h^*\}}_{\text{unlabeled}}. \quad (5)$$

3.3 Robust Representation Learning

Given the re-organized training data $\mathfrak{D}^{\diamond}$ from Eq. 5, we train our model f_Θ using a set of losses to learn robust representations. We adopt three supervised losses: a mixup-based cross-entropy loss (Zhang et al. 2018), a supervised contrastive loss (Khosla et al. 2020), and a noisy-aware balanced cross-entropy loss adapted from the balanced softmax cross-entropy loss (Ren et al. 2020) to be noise-aware. The three supervised losses extract knowledge from the strongly and weakly labeled sets to learn robust representations. Additionally, we incorporate two unsupervised losses, *i.e.* instance-wise contrastive loss (Chen et al. 2020) and cross-view consistency loss (Tarvainen and Valpola 2017), to distill knowledge from all three sets for enhancing robust representation learning.

Noisy-aware balanced cross-entropy loss. Balanced softmax cross-entropy loss (Ren et al. 2020) takes the label frequency as the prior knowledge to adjust the model predictions during training, which is able to alleviate the class-imbalanced bias. However, this loss cannot be directly used in our studied setting due to the existence of noise in the training set, which makes the computed label frequency unreliable. We adapt this loss to the noisy and imbalanced setting by a simple yet effective modification: instead of computing label frequency on the original training set $\mathfrak{D}$, we compute the frequency based on $\mathfrak{D}_s$ and $\mathfrak{D}_w$. We name the adapted loss as *noisy-aware* balanced cross-entropy loss, denoted as $\mathcal{L}_{bsce}$, because $\mathfrak{D}_s$ and $\mathfrak{D}_w$ are dynamically changing with the training epochs in regard to the noisy label correction and open-set noisy rejection.

Formally, let $\mathcal{X}_l$ denote a mini-batch of images from $\mathfrak{D}_s \cup \mathfrak{D}_w$, and $\mathcal{I}_l$ be the index of $\mathcal{X}_l$. The noisy-aware balanced cross-entropy loss is computed as:

$$\mathcal{L}_{bsce} = -\frac{1}{|\mathcal{X}_l|}\sum_{i\in\mathcal{I}_l}\log\left\{\frac{\bar{n}^{y_i}\exp\{\mathbf{l}_i(y_i)\}}{\sum_{k=1}^{C}\bar{n}^{k}\exp\{\mathbf{l}_i(k)\}}\right\}, \quad (6)$$

where $\mathbf{l}_i \in \mathbb{R}^C$ is the predicted logits before softmax. $\bar{n}^k$ is the frequency of k-th class in $\mathfrak{D}_s \cup \mathfrak{D}_w$.

Mixup-based cross-entropy loss. As shown in many prior works (Li, Socher, and Hoi 2020; Li, Xiong, and Hoi 2021; Sachdeva et al. 2021; Wu et al. 2021), mixup (Zhang et al. 2018) is a powerful technique against noisy labels. We adopt this technique to generate more virtual samples $\{\mathbf{x}'_i, y'_i\}$ from $\mathfrak{D}_s$ and $\mathfrak{D}_w$. Subsequently, the cross-entropy loss is applied on these mixed virtual samples, denoted as $\mathcal{L}_{mixup}$.

Supervised contrastive loss. We adopt supervised contrastive learning (Khosla et al. 2020) to effectively leverage label information in $\mathfrak{D}_s$ and $\mathfrak{D}_w$. We use data augmentation techniques to augment $\mathcal{X}_l$, resulting in $\mathcal{X}'_l$, which is indexed by $\mathcal{I}'_l$. The supervised contrastive loss is computed as:

$$\mathcal{L}_{supcon} = -\frac{1}{|\mathcal{X}_l|}\sum_{i\in\mathcal{I}_l}\frac{1}{|\mathcal{S}_i|}\sum_{s\in\mathcal{S}_i}\log\frac{\exp(\mathbf{f}_i\cdot\mathbf{f}_s/\tau)}{\sum_{a\in\mathcal{A}_l(i)}\exp(\mathbf{f}_i\cdot\mathbf{f}_a/\tau)}. \quad (7)$$

Here $\mathcal{A}_l(i) = \{\mathcal{I}_l \setminus \{i\}\} \cup \mathcal{I}'_l$, and $\mathcal{S}_i = \{s \in \mathcal{A}_l(i) : y^*_i = y^*_s\}$, where y^* is either $\hat{y}$ or $\tilde{y}$ according to the set that it comes from. $\tau \in \mathbb{R}^+$ is a temperature parameter.

Instance-wise contrastive loss. Motivated by the potential of contrastive learning in learning with noisy data shown in the prior works (Li, Xiong, and Hoi 2021; Wu et al. 2021), we employ the instance-wise contrastive loss $\mathcal{L}_{inscon}$ over all the training samples.

Cross-view consistency loss. In addition to $\mathcal{L}_{inscon}$ that is applied on the feature space, we also apply a consistency loss on the model predictions over all the samples as a further regularization. Specifically, we maintain an exponential moving average of the current model weights Θ_t: $\tilde{\Theta}_t = \omega\tilde{\Theta}_{t-1} + (1-\omega)\Theta_t$, where $\tilde{\Theta}$ is known as the teacher in mean-teacher framework (Tarvainen and Valpola 2017), and ω is a smoothing coefficient. Let $\mathcal{X}$ be a mini-batch of images from $\mathfrak{D}_\diamond$ and $\mathcal{X}'$ be the augmented version (*i.e.* another view) of $\mathcal{X}$. The two views of the same data ($\mathcal{X}'$ and $\mathcal{X}$) are respectively passed to Θ and $\tilde{\Theta}$, outputting two sets of categorical probabilities $\{\mathbf{o}_i\}$ and $\{\tilde{\mathbf{o}}_i\}$. The cross-view consistency loss is therefore computed as the KL divergence between two predictions: $\mathcal{L}_{consist} = \frac{1}{|\mathcal{X}|}\sum_i D_{KL}(\mathbf{o}_i \parallel \tilde{\mathbf{o}}_i)$. Finally, the total loss of one mini-batch is computed as:

$$\mathcal{L}_{total} = \mathcal{L}_{bsce} + \mathcal{L}_{mixup} + \mathcal{L}_{supcon} + \mathcal{L}_{inscon} + \mathcal{L}_{consist}. \quad (8)$$

4 Experiments

We evaluate our proposed method on three datasets, including CIFAR-10 and CIFAR-100 (Krizhevsky, Hinton et al. 2009) with controlled noise and class imbalance, and WebVision (Li et al. 2017) that is a real-world class-imbalanced dataset with open-world noise. The experimental results on both synthetic and real-world datasets show that our method can improve the testing performance with the co-existence of various noise types and imbalanced classes in the training data.

4.1 Comparison on Synthetic Datasets

Dataset Setup. To the best of our knowledge, we are the first to explicitly study the problem of Learning with class-Imbalanced Open-world Noisy Data (LIOND). Thus, we simulate the setting by manipulating the training data of CIFAR-10 and CIFAR-100 with controlled noise and class imbalance. Specifically, we first follow (Cui et al. 2019) to create the class-imbalanced version of CIFAR [1] by reducing the original number of training samples n^o_k *w.r.t.* an exponential function $n_k = n^o_k\mu^{k/(C-1)}$, where μ denotes the imbalance factor of a dataset (the number of training samples in the largest class divided by that of the smallest) and k is the class index starting from 0. Subsequently, we corrupt the imbalanced training data of CIFAR with *symmetric* closed-set noise added by randomly selecting a portion κ of samples and flipping their real labels to the random class labels from the training classes, following ProtoMix and NGC. We further inject a number of images from open-set datasets, *i.e.* TinyImageNet (Le and Yang 2015) and Places-365 (Zhou et al. 2017), that do not share label space with the evaluated dataset. Note that we also use CIFAR-100 as the open-set dataset for CIFAR-10. The number of open-set images to add is controlled by the ratio ρ, *w.r.t.* the number of current training samples. Each open-set image is assigned with a random class label from the training classes.

[1] We drop the suffix from CIFAR-10 and CIFAR-100 for brevity.

Method \ Open-set	CIFAR-100		TinyImageNet		Places-365	
	$\rho = 20\%$	$\rho = 40\%$	$\rho = 20\%$	$\rho = 40\%$	$\rho = 20\%$	$\rho = 40\%$
Jo-SRC	68.49±0.79	67.51±0.78	68.17±0.76	69.65±1.08	67.11±1.03	68.11±1.76
ProtoMix	72.68±4.29	72.83±6.12	76.50± 8.12	72.75±5.20	71.81±4.91	70.76±5.43
NGC	85.89±0.41	84.86±0.40	86.12±0.60	85.05±1.08	86.15±0.28	85.27±0.29
Ours	**86.29**±0.42	**85.47**±0.40	**86.78**±0.79	**86.03**±0.27	**86.66**±0.38	**85.79**±0.10

Table 1: LIOND performance comparison with baselines on CIFAR-10 dataset with imbalance factor $\mu = 10$, symmetric closed-set noise $\kappa = 50\%$, and different variants of open-set noise. Top-1 accuracy (%) over 3 independent runs is reported.

Method \ Open-set	TinyImageNet		Places-365	
	$\rho = 20\%$	$\rho = 40\%$	$\rho = 20\%$	$\rho = 40\%$
Jo-SRC	33.60±0.17	33.36±0.45	33.16±1.29	34.00±0.41
ProtoMix	42.84±0.95	41.01±0.67	42.95±0.49	42.18±0.69
NGC	54.25±0.48	51.81±1.43	53.89±0.88	52.19±0.79
Ours	**56.17**±0.35	**56.47**±0.63	**56.03**±0.68	**56.77**±0.24

Table 2: LIOND performance comparison with baselines on CIFAR-100 under the same settings of μ and κ as in Tab. 1.

Baselines. We adopt three latest works on Learning with Open-world Noisy Data (LOND): Jo-SRC, ProtoMix [2], and NGC as our baselines. Since these baselines do not consider class imbalance in their experiments on CIFAR datasets, we produce the results under our studied settings based on their officially provided codes.

Implementation. Following ProtoMix and NGC, we use PreAct ResNet-18 (He et al. 2016) as our feature encoder for both CIFAR datasets. The encoder is followed by two parallel linear layers: one projection layer that produces 64-dimensional feature vector $\mathbf{f}_i$ and one classification layer that produces C-dimensional logits $\mathbf{l}_i$. We train the model using SGD optimizer with momentum 0.9 and weight decay 5e-4. We set the batch size as 128 and the initial learning rate as 0.02 with a cosine decay schedule. The model is trained for 300 epochs with a warmup period using $\mathcal{L}_{bsce}$. The warmup period is set to 10 and 30 epochs for CIFAR-10 and CIFAR-100, respectively. We adopt random crop and horizontal flip as weak augmentation to generate $\mathcal{X}$, and AugMix (Hendrycks et al. 2020) as strong augmentation to generate $\mathcal{X}'$. We set the hyper-parameters as $\alpha = 0.05$, $\beta = 3$, $\gamma = 0.5$, $K = 30$, $\epsilon_l = 1$, $\epsilon_h = 1$, $\tau = 0.3$, and $\omega = 0.99$.

Learning with Class-Imbalanced Open-World Noisy Data. Tab. 1 and 2 report the comparison results with the three baselines on CIFAR-10 and CIFAR-100, respectively, under the novel setting of learning with class-imbalanced open-world noisy data. We can see that our method achieves the state-of-the-art performance on all settings. Particularly, our method significantly outperforms the strongest baseline NGC on the CIFAR-100 dataset, which is much more challenging than CIFAR-10 as the number of classes is ten times more (*i.e.* 100 vs. 10) but the number of samples per classes is much lower (*e.g.* 500 vs. 50 for the last tail class). Our method's superior and consistent outperformance demonstrates its effectiveness.

Learning with Open-world Noisy Data. Although our method is proposed for learning with class-imbalanced open-world noisy data, it also shows notable performance under the setting without class imbalance, *i.e.* LOND. The comparison results with the state-of-the-art methods on this setting are shown in Tab. 3. Our method outperforms the previous SOTA by clear margins, in contrast to the marginal performance gaps between ProtoMix and NGC. This underscores the robustness of our method in addressing open-world noise.

[2]For a fair comparison, we adopt ProtoMix (classifier) that uses the softmax classifier instead of the KNN classifier as our baseline.

4.2 Comparison on Real-world Dataset

In addition to the verification on the simulating datasets with controlled noise and class imbalance, we evaluate our method on the real-world dataset - WebVision (Li et al. 2017). WebVision uses the same categories from ImageNet ILSVRC12 (Deng et al. 2009) to crawl images from Flickr and Google. As a result, the images in WebVision inherently contains (both closed-set and open-set) noisy labels and imbalanced classes. Following previous works, we conduct experiments on the first 50 classes.

Implementation. To align with ProtoMix and NGC, we adopt Inception-ResNet V2 as the feature encoder. We train the model using SGD optimizer with momentum 0.9 and weight decay 1e-4. We set the batch size as 32 and the initial learning rate as 0.04 with a cosine decay schedule. The model is trained for 80 epochs with a 15-epoch warmup period using $\mathcal{L}_{bsce}$. The configuration of hyper-parameters is the same as that for CIFAR datasets, except for $K = 50$ and $\epsilon_l = 0.1$.

Results. The comparison results are reported in Tab. 4. It is obvious that our method outperforms the competing methods on the two validation sets by significant margins. Our state-of-the-art performance demonstrates the strength of our method in the real-world scenario.

4.3 Ablation Studies

To investigate the effects of our design choice and hyper-parameters in LIOND, we conduct ablation studies on CIFAR-100 dataset with imbalance factor $\mu = 10$, symmetric closed-set noise $\kappa = 50\%$, and $\rho = 40\%$ open-set samples from TinyImageNet.

Method \ Dataset	CIFAR-10			CIFAR-100	
	CIFAR-100	TinyImageNet	Places-365	TinyImageNet	Places-365
Jo-SRC	84.46±0.22	84.89±0.53	84.74±0.29	59.87±0.26	60.24±0.64
ProtoMix	92.17±0.75	93.15±0.64	93.95±0.07	73.14±0.58	72.31±0.10
NGC†	92.31±0.29	93.54±0.21	93.67±0.22	73.49±0.11	73.44±0.35
NGC	92.40±0.39	93.78±0.17	93.89±0.18	73.53±0.27	73.84±0.07
Ours	**93.50**±0.17	**94.04**±0.22	**94.24**±0.22	**75.35**±0.39	**75.30**±0.51

Table 3: LOND performance comparison with baselines on CIFAR-10 and CIFAR-100 datasets with symmetric closed-set noise $\kappa = 50\%$ and open-set noise $\rho = 40\%$ from different open-set datasets. † indicates the reported results in the original paper.

Method \ Dataset	WebVision-50		ILSVRC12	
	top1	top5	top1	top5
INCV†	65.24	85.34	61.60	84.98
DivideMix†	77.32	91.64	75.20	90.84
ELR+†	77.78	91.68	70.29	89.76
ProtoMix†	76.30	91.50	73.30	91.20
NGC†	79.16	91.84	74.44	91.04
Jo-SRC	74.40	90.64	69.68	88.44
ProtoMix	77.08	90.76	72.56	90.24
NGC	78.92	92.24	75.00	91.00
Ours	**80.80**	**93.00**	**77.24**	**92.28**

Table 4: Performance comparison with baseline methods on WebVision-50 and ILSVRC12 validation sets, using WebVision-50 for training. † denoted the reported results in the paper. We reproduce the results for ProtoMix and NGC with the same optimization setting as our method.

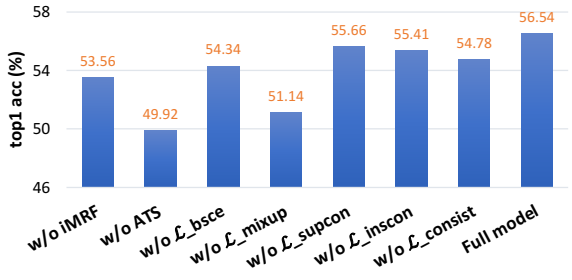

Figure 2: Ablation study of different components of our method on CIFAR-100 in LIOND.

Design Choice. We first study the effects of the proposed iMRF, agreement-based thresholding strategy (ATS), and different loss functions. Fig. 2 shows the resultant performance with the removal of each component. Specifically, 'w/o iMRF' directly uses the network predictions as the posterior distribution $\mathbf{p}_i$. 'w/o ATS' replaces the proposed ATS with the fixed thresholding strategy of ProtoMix (*i.e.* $\eta_l = 0.01$ and $\eta_h = 0.9$). 'w/o $\mathcal{L}_*$' removes the corresponding loss from the total loss in Eq. 8, except for 'w/o $\mathcal{L}_{bsce}$'. We replace $\mathcal{L}_{bsce}$ with conventional cross-entropy loss since $\mathcal{L}_{bsce}$ is compulsory for the warmup period. As shown in the figure, ATS contributes most to the final performance, which empirically testifies our hypothesis on the importance of adaptive thresholding. Furthermore, the performance also

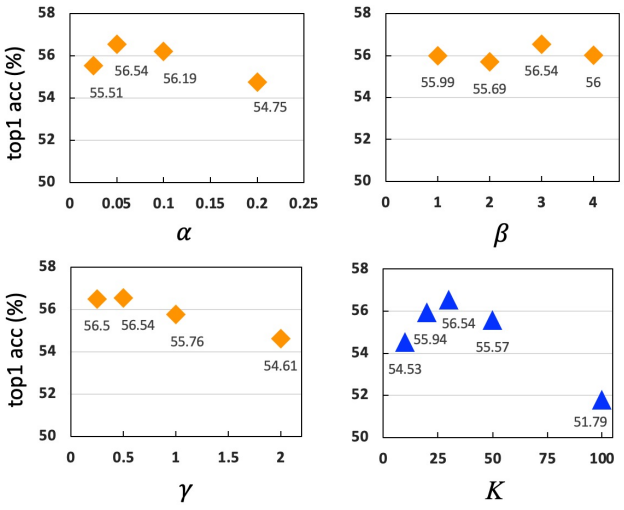

Figure 3: Effects of different hyper-parameters (*i.e.* α, β, γ, and K) on CIFAR-100 in LIOND.

suffers from a significant drop when removing iMRF. It is worth mentioning that the drastic performance drop with the removal of mixup is also observed in ProtoMix and NGC.

Hyper-parameter Tuning. Fig. 3 shows the impact of four hyperparameters. A large K (e.g., K=100) harms performance, likely because of the scarcity of samples for tail classes in CIFAR-100, resulting in noisy labels for the last tail class with only 50 samples, making it unable to form a valid K nearest neighborhood. Generally, hyperparameters are less sensitive within a certain range.

5 Conclusion

This paper explores robust visual recognition with class-imbalanced open-world noisy data. We employ an iterative learning strategy involving label correction through sample selection and robust representation learning. Our probabilistic graphical model-based approach, iMRF, effectively corrects label noise. Additionally, we introduce a tripartite agreement-based thresholding strategy for dynamic threshold adjustment during sample selection. We propose a noisy-aware balanced cross-entropy loss, integrated with carefully selected losses, to achieve a robust model. Extensive experiments on synthetic and real datasets with diverse noisy and imbalanced settings demonstrate consistent and superior performance improvements over existing approaches.

References

Andrieu, C.; De Freitas, N.; Doucet, A.; and Jordan, M. I. 2003. An introduction to MCMC for machine learning. *Machine learning*, 50(1): 5–43.

Boykov, Y.; Veksler, O.; and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11): 1222–1239.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*, 9268–9277.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*, 31.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *ECCV*, 630–645.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *ICLR*.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *ICCV*, 3326–3334.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2019. Label propagation for deep semi-supervised learning. In *CVPR*, 5070–5079.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.-J.; and Fei-Fei, L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2304–2313.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2): 183–233.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling Representation and Classifier for Long-Tailed Recognition. In *ICLR*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 33: 18661–18673.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *ICLR*.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to learn from noisy labeled data. In *CVPR*, 5051–5059.

Li, J.; Xiong, C.; and Hoi, S. C. 2021. Learning from noisy data with robust representation learning. In *ICCV*, 9485–9494.

Li, W.; Wang, L.; Li, W.; Agustsson, E.; and Van Gool, L. 2017. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Liu, J.; Sun, Y.; Han, C.; Dou, Z.; and Li, W. 2020a. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2970–2979.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020b. Early-learning regularization prevents memorization of noisy labels. *NeurIPS*, 33: 20331–20342.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, 1944–1952.

Reed, S.; Lee, H.; Anguelov, D.; Szegedy, C.; Erhan, D.; and Rabinovich, A. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Reed, W. J. 2001. The Pareto, Zipf and other power laws. *Economics letters*, 74(1): 15–19.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33: 4175–4186.

Sachdeva, R.; Cordeiro, F. R.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2021. Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *WACV*, 3607–3615.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *TNNLS*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 30.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *CVPR*, 8769–8778.

Wang, Y.; Liu, W.; Ma, X.; Bailey, J.; Zha, H.; Song, L.; and Xia, S.-T. 2018. Iterative learning with open-set noisy labels. In *CVPR*, 8688–8696.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR*, 13726–13735.

Wei, H.; Tao, L.; Xie, R.; and An, B. 2021. Open-set label noise can improve robustness against inherent label noise. *NeurIPS*, 34.

Wu, Z.-F.; Wei, T.; Jiang, J.; Mao, C.; Tang, M.; and Li, Y.-F. 2021. Ngc: A unified framework for learning with open-world noisy data. In *ICCV*, 62–71.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2020. Robust early-learning: Hindering the memorization of noisy labels. In *ICLR*.

Xiang, L.; Ding, G.; and Han, J. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, 247–263. Springer.

Yao, Y.; Liu, T.; Han, B.; Gong, M.; Deng, J.; Niu, G.; and Sugiyama, M. 2020. Dual t: Reducing estimation error for transition matrix in label-noise learning. *NeurIPS*, 33: 7260–7271.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-src: A contrastive approach for combating noisy labels. In *CVPR*, 5192–5201.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, H.; Xing, X.; and Liu, L. 2021. DualGraph: A graph-based method for reasoning about label noise. In *CVPR*, 9654–9663.

Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; and Feng, J. 2021a. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*.

Zhang, Y.; Zheng, S.; Wu, P.; Goswami, M.; and Chen, C. 2021b. Learning with Feature-Dependent Label Noise: A Progressive Approach. In *ICLR*.

Zhang, Z.; and Sabuncu, M. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *NeurIPS*, 31.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 9719–9728.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *PAMI*.

Zhou, D.; Bousquet, O.; Lal, T.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. *NeurIPS*, 16.