Regroup Median Loss for Combating Label Noise

Fengpeng Li[1], **Kemou Li**[1], **Jinyu Tian**[2], **Jiantao Zhou**[1*]

[1]State Key Laboratory of Internet of Things for Smart City,
Department of Computer and Information Science, University of Macau
[2]Faculty of Innovation Engineering, Macau University of Science and Technology
{yc17420, mc15415, jtzhou}@um.edu.mo, jytian@must.edu.mo

Abstract

The deep model training procedure requires large-scale datasets of annotated data. Due to the difficulty of annotating a large number of samples, label noise caused by incorrect annotations is inevitable, resulting in low model performance and poor model generalization. To combat label noise, current methods usually select clean samples based on the small-loss criterion and use these samples for training. Due to some noisy samples similar to clean ones, these small-loss criterion-based methods are still affected by label noise. To address this issue, in this work, we propose Regroup Median Loss (RML) to reduce the probability of selecting noisy samples and correct losses of noisy samples. RML randomly selects samples with the same label as the training samples based on a new loss processing method. Then, we combine the stable mean loss and the robust median loss through a proposed regrouping strategy to obtain robust loss estimation for noisy samples. To further improve the model performance against label noise, we propose a new sample selection strategy and build a semi-supervised method based on RML. Compared to state-of-the-art methods, for both the traditionally trained and semi-supervised models, RML achieves a significant improvement on synthetic and complex real-world datasets. The source code is available at https://github.com/Feng-peng-Li/Regroup-Loss-Median-to-Combat-Label-Noise.

Introduction

Deep learning model has been proven powerful in various practical tasks, *e.g.*, image classification (Szegedy et al. 2015; Zhang et al. 2015; Krizhevsky, Sutskever, and Hinton 2017), object detection (Girshick 2015; Redmon et al. 2016) and image semantic segmentation (He et al. 2017; Zhang et al. 2018). As is known, the performance of deep learning models heavily relies on dataset scale and annotation quality. Currently, collecting large-scale datasets with high-quality annotation is an extremely expensive task. An efficient and cheap method called crowd-sourcing policy is often used to collect large-scale datasets (Zubiaga et al. 2015). However, in the crowd-sourcing data labeling procedure, it is inevitable that some samples will be annotated with incorrect labels, resulting in the so-called label noise. Due to the strong capacity of the deep learning model, it can easily fit the training samples with incorrect labels, which impairs the performance and generalization of the deep model (Zhang et al. 2017). Owing to the difficulty of selecting noisy samples in a large-scale dataset, how to make the deep model robust to label noise becomes an essential and fundamental research topic (Bai et al. 2021; Karim et al. 2022).

To protect the deep model from label noise (Han et al. 2018; Bai et al. 2021; Karim et al. 2022), some methods have been proposed recently. According to the operations of these methods, they can be divided into two categories: loss correction and sample selection (Karim et al. 2022). Loss correction commonly attempts to estimate the noise transition matrix. The matrix contains the conditional probabilities of samples belonging to different classes based on their observed labels (Frénay and Verleysen 2013). However, estimating the transition matrix is difficult especially when the number of classes is large and the ratio of noisy samples is too high (Xu et al. 2019; Karim et al. 2022). Compared to loss correction, sample selection methods pay more attention to filtering out samples with incorrect labels based on the small-loss criterion (Han et al. 2018; Yu et al. 2019), supposing clean samples have smaller losses than noisy ones. Inevitably, some noisy samples will also have small loss values, and will therefore be misclassified as clean. Traditional small-loss criterion based methods usually select only clean samples for training and discard noisy ones (Han et al. 2018; Xia et al. 2019), resulting in information loss. The semi-supervised methods (Li, Socher, and Hoi 2020; Karim et al. 2022) use the small-loss criterion to select clean samples and relabel noisy ones. However, these methods require complex strategies to separate clean and noisy samples, which increases time consumption.

According to (Xia et al. 2020), noisy samples can also provide useful information for model training if their losses can be appropriately corrected. Motivated by these previous works, we propose Regroup Median Loss (RML) to reduce the probability of selecting noisy samples and correct the distorted losses based on losses of clean samples through a robust estimation method. In RML, we select samples with the same observed label as each training sample and use their losses to estimate the loss of the training sample. To reduce the probability of selecting noisy samples, we propose a new strategy to process losses and use them to select

*Corresponding author.

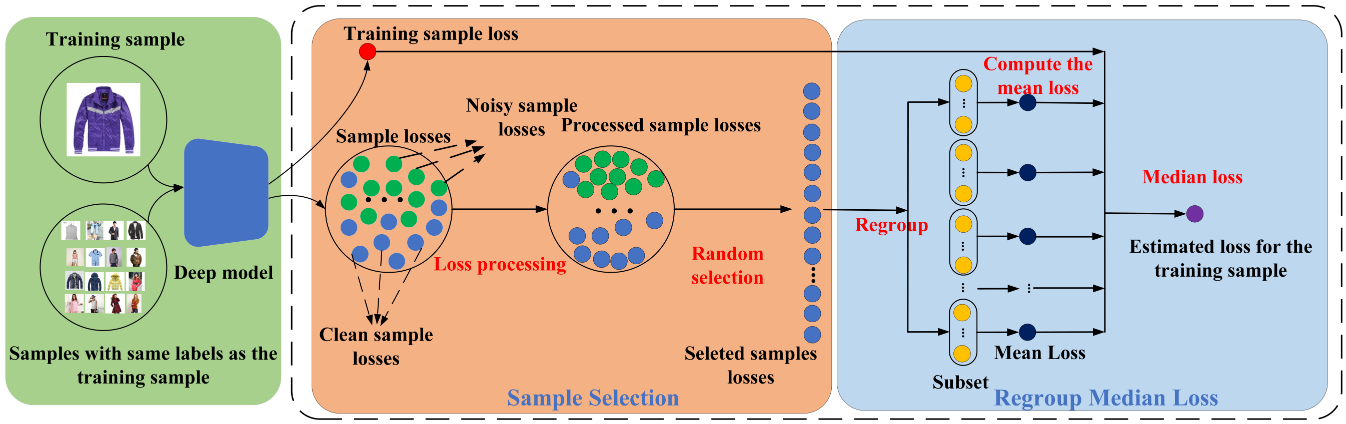

Figure 1: The overview of RML, which consists of two steps: (i) Sample Selection and (ii) Regroup Median Loss.

samples. However, noisy samples may still be selected inevitably. Loss estimation by the mean loss of all selected samples will be influenced. According to (Nemirovskij and Yudin 1985; Catoni 2012), Median-of-Means (MoM) approach is particularly suitable for heavy-tailed distributions and is insensitive to outliers. Motivated by MoM, we propose a robust loss estimation method that combines median loss and mean loss. During the loss estimation procedure, our proposed RML divides the selected samples into several groups and computes the mean loss of each group. Then, we compute the median between these mean losses and the training sample loss to robustly estimate the training sample loss. In addition, we build a semi-supervised method based on RML to combat label noise.

The contributions of this paper are presented as follows:

- We propose a loss processing method to reduce the probability of selecting noisy samples. We theoretically explain and verify the operation that reduces the probability of selecting noisy samples.
- We propose Regroup Median Loss, a novel robust loss estimation method that can correct the distorted loss value of noisy samples. Then these corrected losses of training samples can mitigate the negative impact of noisy samples and improve the model performance.
- We verify that RML can achieve better model performance than existing methods on synthetic and real-world datasets with various types of label noises. The proposed method achieves about 6% performance improvement on CIFAR-100 with various types of label noises and increases the test accuracy by about 8% on the challenging real-world dataset WebVision. Moreover, the semi-supervised method based on RML achieves state-of-the-art performance on synthetic and real-world datasets.

Related Works

This section summarizes existing methods to combat label noise. To better understand the motivation of existing methods, we first introduce the types of label noises. According to (Song et al. 2020), label noise can be divided into two categories: instance-independent label noise and instance-dependent label noise. For instance-independent label noise, it is only relevant to the original sample labels and independent of the sample features (Van Rooyen, Menon, and Williamson 2015). Symmetric label noise and asymmetric label noise are two typical types of instance-independent label noise. Compared to instance-independent label noise, the generating corruption probability of instance-dependent label noise depends on both sample features and labels (Xia et al. 2020). Some studies show that instance-dependent label noise is common in real-world datasets, such as Clothing1M (Xiao et al. 2015) and WebVision (Li et al. 2017).

Due to the negative impact of label noise on model generalization, a large number of methods have been proposed (Han et al. 2018; Yu et al. 2019; Berthelot et al. 2019; Bai et al. 2021; Karim et al. 2022). They can be roughly divided into two main species: loss correction and sample selection. Loss correction adopts the noise transition matrix to weigh label smoothing parameters (Xia et al. 2021; Cheng et al. 2021). The difficulty of this method is how to obtain the noise transition matrix. One commonly used method tends to count the transition relationship between clean dataset labels and noisy dataset labels (Xia et al. 2019; Cheng et al. 2021; Zhu, Liu, and Liu 2021) and use the label change frequency as the transition probability to build the true noise transition matrix. The other method attempts to use some constraints or characteristics of the noise transition matrix, such as total variation (Zhang, Niu, and Sugiyama 2021) and diagonal dominance (Li et al. 2021) to obtain a transition matrix by gradient descent. Although the matrix obtained by the former method has a remarkable performance against label noise than the calculated transition matrix, it is difficult to obtain the original clean dataset in a practical task (Karim et al. 2022). In addition, label smoothing (Müller, Kornblith, and Hinton 2019) is also a way to achieve loss correction for noisy data (Zhang et al. 2021; Cui et al. 2022).

Sample selection is another type of method that aims to train models on small-loss samples. Intuitively, small-loss samples are more likely to be correctly labeled. The memorization effect of DNNs shows that even with noisy labels, DNNs learn clean and simple patterns first, and then gradually fit all samples (Arpit et al. 2017). This has given rise to the widely used small-loss criterion: considering small-loss samples as clean ones. Co-teaching (Han et al. 2018) is a typical method based on the small-loss criterion with two deep models. However, the selection operation based on the small-loss criterion can still make mistakes, misclassifying noisy samples as clean ones, and vice versa. To further address the shortcomings of the traditional small-loss criterion based methods, a semi-supervised method, DivideMix (Li, Socher, and Hoi 2020), adopts Gaussian Model Mixture (Rasmussen 1999) method to select clean samples based on sample losses and uses MixMatch (Berthelot et al. 2019) to train two deep models. Although DivideMix has remarkable performance against label noise, it requires the noise rate of the dataset, which significantly limits its application. To further address the issue of DivideMix, some methods, such as MOIT (Ortego et al. 2021), Jo-SRC (Yao et al. 2021) and UNICON (Karim et al. 2022), develop new strategies to separate clean and noisy samples.

Methods

In this section, we present the details of our proposed approach for combating label noise. We start by introducing some fundamental notations, followed by a two-part description of the RML. We then analyze the robustness of RML theoretically. Lastly, we introduce a common method and a semi-supervised model based on RML.

Notations

We use bold capital letters such as $\boldsymbol{X}$ to represent a random vector, bold lowercase letters such as $\boldsymbol{x}$ to represent the realization of a random vector, capital letters such as Y to represent a random variable, and lowercase letters such as y to represent the realization of a random variable. Consider a c-class classification problem, let $\mathcal{X}$ be the feature space, $\mathcal{Y} = \{1, \dots, c\}$ be the label space, $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ be the random variables with joint distribution $P_{\boldsymbol{X},Y}$ and $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ be a dataset containing i.i.d. N samples drawn from $P_{\boldsymbol{X},Y}$, where $\boldsymbol{x}_i$ and y_i are the i-th instance and its label. In practical applications, the true label Y may not be observable. Instead, we have an noisy dataset $\tilde{\mathcal{D}} = \{(\boldsymbol{x}_i, \tilde{y}_i)\}_{i=1}^N$ consisting of i.i.d. N samples drawn from $P_{\boldsymbol{X},\tilde{Y}}$, where $\boldsymbol{x}_i$ is the i-th instance and $\tilde{y}_i$ is its observed label which may be correct or not. Define the loss set of $\tilde{\mathcal{D}}$ as $\mathcal{L}(\tilde{\mathcal{D}}) := \left\{\ell(f(\boldsymbol{x}_i), \tilde{y}_i) \mid (\boldsymbol{x}_i, \tilde{y}_i) \in \tilde{\mathcal{D}}\right\}_{i=1}^N$, where $\ell(\cdot)$ denotes the cross-entropy (CE) loss function and f is a deep model with parameter $\theta \in \mathbb{R}^d$.

Regroup Median Loss

Fig. 1 illustrates the two-step process of RML. In Step (i), a newly-proposed loss processing method is used to select clean samples. Step (ii) utilizes losses of the selected samples and the original sample to estimate the loss by a regroup strategy. Next we provide more details on these two steps.

Sample Selection. In the process of RML, first of all, for each training sample, we select samples with the same

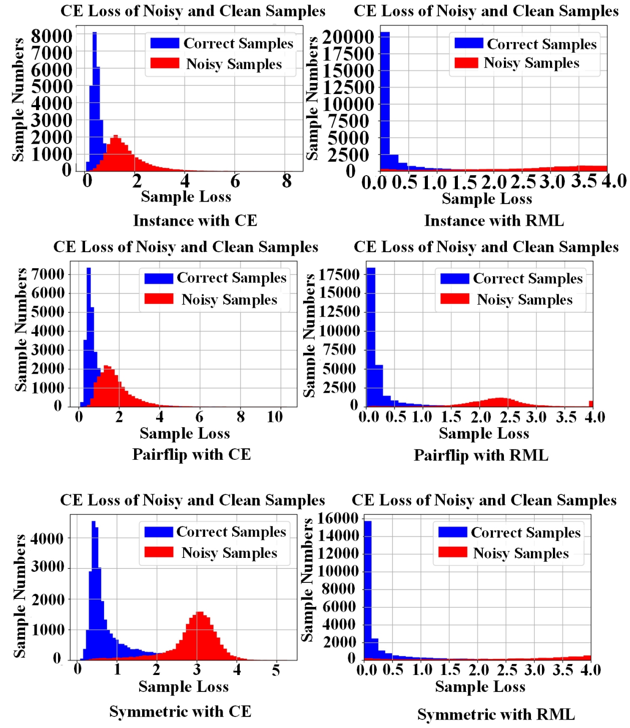

Figure 2: Loss distribution of noisy and clean samples on CIFAR-10 for the model trained with CE and RML.

observed labels as this given sample. Consider a training sample $(\boldsymbol{x}, \tilde{y})$ in $\tilde{\mathcal{D}}$, suppose that there are totally m samples with the same label as $\tilde{y}$ in $\tilde{\mathcal{D}}$. We denote by $\tilde{\mathcal{D}}_{\tilde{y}} = \left\{(\boldsymbol{x}_i, \tilde{y}_i) \in \tilde{\mathcal{D}} \mid \tilde{y}_i = \tilde{y}\right\}_{i=1}^m = \{(\boldsymbol{x}_i, \tilde{y})\}_{i=1}^m$ the selected set and $\mathcal{L}(\tilde{\mathcal{D}}_{\tilde{y}}) = \left\{\ell(f(\boldsymbol{x}_i), \tilde{y}) \mid (\boldsymbol{x}_i, \tilde{y}) \in \tilde{\mathcal{D}}_{\tilde{y}}\right\}_{i=1}^m = \{\ell_i\}_{i=1}^m$ the loss set of $\tilde{\mathcal{D}}_{\tilde{y}}$. Since clean samples usually have small losses (Gui, Wang, and Tian 2021), we define a new sample selection probability based on sample loss to select samples with small losses from $\tilde{\mathcal{D}}_{\tilde{y}}$.

Definition 1 *For any* $(\boldsymbol{x}_i, \tilde{y}_i) \in \tilde{\mathcal{D}}_{\tilde{y}}$*, define its selection probability* p_i *based on its original CE loss* ℓ_i *as*

$$p_i := \frac{\mathrm{e}^{-\ell_i}}{\sum_{j=1}^m \mathrm{e}^{-\ell_j}}. \tag{1}$$

By Def. 1, the larger the sample loss, the smaller its selection probability, *i.e.*, clean samples have a higher selection probability. Nevertheless, such sample selection based on original CE loss still has the same issue as the current small-loss criterion because of the loss overlap between the clean and noisy samples (Bai et al. 2021). To better separate noise and clean samples, we processed the sample loss.

Definition 2 *For any* $(\boldsymbol{x}_i, \tilde{y}_i) \in \tilde{\mathcal{D}}_{\tilde{y}}$*, define its processed loss* $\tilde{\ell}_i$ *based on its original CE loss* ℓ_i *as*

$$\tilde{\ell}_i := \ell_i \times (\ell_i + \varepsilon), \tag{2}$$

where ε is a bias term. Thus, its improved selection probability $\tilde{p}_i$ based on its processed loss $\tilde{\ell}_i$ is

$$\tilde{p}_i := \frac{\mathrm{e}^{-\tilde{\ell}_i}}{\sum_{j=1}^{m} \mathrm{e}^{-\tilde{\ell}_j}}. \quad (3)$$

Then, we employ Prop. 1 to explain the operation.

Proposition 1 *For an arbitrary sample $(\boldsymbol{x}_\tau, \tilde{y})$ in $\tilde{\mathcal{D}}_{\tilde{y}}$, after the processing operation in Eq. (2), its selection probability change is $\log p_\tau - \log \tilde{p}_\tau = \ell_\tau(\ell_\tau + \varepsilon - 1) - \beta = \ell_\tau^2 - \beta$, where $\beta = \log \frac{\sum_{j=1}^{m} \mathrm{e}^{-\ell_j}}{\sum_{j=1}^{m} \mathrm{e}^{-\ell_j(\ell_j+\varepsilon)}}$ is a constant.*

Prop. 1 demonstrates the change of selection probability between Eq. (1) and (3). Since $\sum_{j=1}^{m} \mathrm{e}^{-\ell_j(\ell_j+\varepsilon)} = \sum_{j=1}^{m} \mathrm{e}^{-\ell_j^2-\ell_j} < \sum_{j=1}^{m} \mathrm{e}^{-\ell_j}$, then $\beta > 0$ and has a fixed value for each training epoch. If $\ell_\tau^2 > \beta$, which means the loss value of the sample is large enough, then $\log p_\tau > \log \tilde{p}_\tau$ and thus $\tilde{p}_\tau < p_\tau$, the selection probability is reduced. On the contrary, the selection probability is improved when $\ell_\tau^2 < \beta$, indicating the sample has small loss. Therefore, the processing operation in Eq. (2) reduces the selection probability of noisy samples while increasing that of clean ones. Notably, the setting of ε affects the number of samples with high selection probability. The larger ε is, the fewer samples with high selection probability, resulting in too few samples to select and overfitting issues. Setting $\varepsilon = 1$ can achieve a trade-off between high selection probability of clean samples and diversity of selected samples.

Regroup Median Loss. As stated above, the proposed method reduces the probability of selecting noisy samples. The samples selected by the probability according to Eq. (3) are more likely to have correct labels. To make full use of the information about noisy samples, we then use the selected samples to estimate the loss of training samples.

First of all, for a training sample $(\boldsymbol{x}, \tilde{y})$ in a mini-batch, recall that RML records losses of all samples in $\tilde{\mathcal{D}}_{\tilde{y}}$ as $\mathcal{L}(\tilde{\mathcal{D}}_{\tilde{y}})$. Then we randomly select $n \times k$ samples according to Eq. (3) without replacement, and combine them as $\tilde{\mathcal{D}}_{\tilde{y}}^s = \{(\boldsymbol{x}_i^s, \tilde{y})\}_{i=1}^{n \times k} \subseteq \tilde{\mathcal{D}}_{\tilde{y}}$, where $\boldsymbol{x}_i^s$ is the i-th selected sample, $n \times k \leqslant m$ and n is an even number. $\tilde{\mathcal{D}}_{\tilde{y}}^s$ follows a distribution $P_{\boldsymbol{X}^s|\tilde{Y}=\tilde{y}}$ whose p.d.f. is given by Eq. (3). After the selection operation, these selected samples are used to estimate the training sample loss.

Although the median loss of samples in $\tilde{\mathcal{D}}_{\tilde{y}}$ can mitigate the negative impact of outliers of the loss distribution of selected samples, it is small and has a large fluctuation between different training steps (Staerman et al. 2021), resulting in the difficulty of the model training procedure convergence. Compared to the median loss, the mean loss is stable, although it is easily affected by the distorted loss of noisy samples. To obtain a stable and robust loss estimation, RML combines the mean loss and the median loss. RML partitions $\tilde{\mathcal{D}}_{\tilde{y}}^s$ into n disjoint subsets $\mathcal{S}_1, \ldots, \mathcal{S}_n$ of size $|\mathcal{S}_i| = k$ as $\tilde{\mathcal{D}}_{\tilde{y}}^s = \bigcup_{i=1}^{n} \mathcal{S}_i$. Then we calculate the mean loss of each subset $\mathcal{S}_i$ and store them as $\mathcal{W} = \left\{\ell_i^s \mid \ell_i^s = \frac{1}{|\mathcal{S}_i|} \sum_{(\boldsymbol{x}^s,\tilde{y}) \in \mathcal{S}_i} \ell(f(\boldsymbol{x}^s), \tilde{y})\right\}_{i=1}^{n}$. For a single training sample $(\boldsymbol{x}, \tilde{y})$, RML estimates its loss by the median of $\mathcal{W}$ and its original loss $\ell(f(\boldsymbol{x}), \tilde{y})$ as

$$\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y}) := \mathrm{Median}\big(\mathcal{W} \cup \{\ell(f(\boldsymbol{x}), \tilde{y})\}\big). \quad (4)$$

Intuitively, if the observed label $\tilde{y}$ is the true label of $\boldsymbol{x}$, such an operation ensures that the estimated loss does not deviate too far from its true loss. Specifically, selecting samples randomly makes each subset of $\tilde{\mathcal{D}}_{\tilde{y}}^s$ contains different samples in different training steps, thus protecting the model from overfitting issues caused by the fixed sample combination. Compared to the Median-of-Means, using the mean of the median losses of each subset can also mitigate the negative of selected noisy samples in $\tilde{\mathcal{D}}_{\tilde{y}}^s$. However, when the observed label $\tilde{y}$ is incorrect, $\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y})$ obtained by the mean of median losses of all subsets is distorted. To make the model training procedure with RML efficient, we add some operations, which can be found in Appendix .

Robustness Analysis of RML

This section analyzes the robustness of Eq. (4), which is our main objective. Note that $\ell(f(\boldsymbol{x}), \tilde{y})$ can also be viewed as the mean loss of a set. Thus, we consider a set $\mathcal{S}_{n+1}$ containing k samples drawn from $P_{\boldsymbol{X}^s|\tilde{Y}=\tilde{y}}$ and satisfying $\frac{1}{|\mathcal{S}_{n+1}|} \sum_{(\boldsymbol{x}^s,\tilde{y}) \in \mathcal{S}_{n+1}} \ell(f(\boldsymbol{x}^s), \tilde{y}) = \ell(f(\boldsymbol{x}), \tilde{y})$. According to (Lecué and Lerasle 2020), such RML can be regarded as an MoM estimator of the loss expectation of $\tilde{\mathcal{D}}_{\tilde{y}}^* = \bigcup_{i=1}^{n+1} \mathcal{S}_i$. To introduce the properties of RML, we make some mild assumptions about loss function ℓ on $\tilde{\mathcal{D}}_{\tilde{y}}^*$.

Assumption 1 *Assume that the samples in $\tilde{\mathcal{D}}_{\tilde{y}}^*$ are i.i.d. drawn from a distribution $P_{\boldsymbol{X}^*|\tilde{Y}=\tilde{y}}$, satisfying $\mathbb{E}_{\boldsymbol{X}^*|\tilde{Y}=\tilde{y}}[\ell(f(\boldsymbol{X}^*), \tilde{y})] = \hat{\mu}$ and $\mathrm{Var}_{\boldsymbol{X}^*|\tilde{Y}=\tilde{y}}[\ell(f(\boldsymbol{X}^*), \tilde{y})] = \hat{\sigma}^2 < \infty$.*

Then based on the properties of MoM, we employ Thm. 1 to prove the robustness of ℓ_{RML}. The proof of the theorem is in Appendix .

Theorem 1 *For any $n, \epsilon > 0$, $\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y})$ satisfies $\mathbb{P}\left(|\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y}) - \hat{\mu}| > \epsilon\right) \leqslant \mathrm{e}^{-\mathrm{C}_1(\frac{1}{2} - \mathrm{C}_2 \frac{\hat{\sigma}^2}{\epsilon^2})^2}$, where C_1 and C_2 are positive constants.*

In noisy label learning, demonstrating that a loss function is robust involves showing that the estimated loss is equivalent to the loss calculated using clean labels, *i.e.*, for $(\boldsymbol{x}, \tilde{y})$, $\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y})$ is close to its true loss $\ell(f(\boldsymbol{x}), y)$. However, due to the uncertainty about the correctness of the observed label, it is difficult to know $\ell(f(\boldsymbol{x}), y)$. In the proposed RML, we first save samples with the label $\tilde{y}$ in $\tilde{\mathcal{D}}$ as $\tilde{\mathcal{D}}_{\tilde{y}}$, then select a subset $\tilde{\mathcal{D}}_{\tilde{y}}^s$ of $\tilde{\mathcal{D}}_{\tilde{y}}$ per Eq. (3). After these two steps, the distribution of $\tilde{\mathcal{D}}_{\tilde{y}}^s$ can be approximated as the sample distribution when the true label is $\tilde{y}$, *i.e.*, $P_{\boldsymbol{X}^s|\tilde{Y}=\tilde{y}} \to P_{\boldsymbol{X}|Y=\tilde{y}}$. By combining this with $\ell(f(\boldsymbol{x}), \tilde{y})$, we obtain $\tilde{\mathcal{D}}_{\tilde{y}}^*$ through our analysis. Following this line, the distance between $\hat{\mu}$ and $\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y})$ can be used to evaluate the robustness of RML. Incorporating Thm. 1, it give

Dataset	Method	Symmetric		Pairflip	Instance	
		0.2	0.5	0.45	0.2	0.4
CIFAR-10	CE	84.00±0.66	75.51±1.24	63.34±6.03	85.10±0.68	77.00±2.17
	Co-teaching	87.16±0.11	72.80±0.45	70.11±1.16	86.54±0.11	80.98±0.39
	Forward	85.63±0.52	77.92±0.66	60.15±1.97	85.29±0.38	74.72±3.24
	Joint-Optim	89.70±0.11	85.00±0.17	82.63±1.38	89.69±0.42	82.62±0.57
	T-revision	89.63±0.13	83.40±0.65	77.06±6.47	90.46±0.13	85.37±3.36
	DMI	88.18±0.36	78.28±0.48	57.60±14.56	89.14±0.36	84.78±1.97
	PES	92.38±0.40	87.45±0.35	88.43±1.08	92.69±0.44	89.73±0.51
	RML (Ours)	**92.89±0.42**	**89.13±0.21**	**90.54±0.34**	**93.34±0.23**	**90.73±0.44**
CIFAR-100	CE	51.43±0.58	37.69±3.45	34.10±2.04	52.19±1.42	42.26±1.29
	Co-teaching	59.28±0.47	41.37±0.08	33.22±0.48	57.24±0.69	45.69±0.99
	Forward	57.75±0.37	44.66±1.01	27.88±0.80	58.76±0.66	44.50±0.72
	Joint-Optim	64.55±0.38	50.22±0.41	42.61±0.61	65.15±0.31	55.57±0.41
	T-revision	65.40±1.07	50.24±1.45	41.10±1.95	60.71±0.73	51.54±0.91
	DMI	58.73±0.70	44.25±1.14	26.90±0.45	58.05±0.20	47.36±0.68
	PES	68.89±0.45	58.90±2.72	57.18±1.44	70.49±0.79	65.68±1.41
	RML (Ours)	**69.74±0.24**	**67.05±0.31**	**65.67±0.54**	**73.43±0.71**	**71.12±1.12**

Table 1: Average test accuracy (%) comparison with state-of-the-art methods without semi-supervised strategy on CIFAR-10 and CIFAR-100. The mean and standard deviation computed over five runs are presented. Baselines are from (Bai et al. 2021).

birth to a conclusion that w.h.p., $\ell_{\mathrm{RML}}(f(\boldsymbol{x}), \tilde{y})$ can converge to the true loss $\ell(f(\boldsymbol{x}), y)$. Further explanation of the theorem is attached to Appendix .

Corollary 1 *After the loss processing operation in Eq.* (2)*, the upper bound in Thm. 1 is reduced.*

Cor. 1 shows that our loss processing operation can make the loss estimation more robust. A simple proof can be found in Rmk. 1 of Appendix .

Combating Label Noise Based on RML

After the introduction of RML, we then present the common and semi-supervised methods for combating label noise. For the RML-based methods with and without the semi-supervised strategy, we set a traditional training model f_θ and a momentum model $g_{\theta'}$ with parameter as $\theta' \in \mathbb{R}^d$ (Tarvainen and Valpola 2017) updated following

$$\theta'_{t+1} = \lambda\theta'_t + (1-\lambda)\theta_{t+1}, \tag{5}$$

where t and λ are the training step and a weighting parameter. In our experiments, λ is set to 0.999 according to (Tarvainen and Valpola 2017). The pseudocode of the RML-based method is described in Alg. 1 of Appendix .

Common Method Based on RML. As shown in Fig. 2, the traditionally trained model divides training samples into two groups based on their loss values. The left column of the figures shows the prediction results of a model trained using the traditional CE loss, while the right column shows the prediction results of a model trained using RML. Compared to the figures in the left column, those in the right column clearly show that clean and noisy samples are separated into two distinct groups. This demonstrates that RML can prevent the model from learning from noisy samples while still extracting sufficient information from clean samples. By using the knowledge gained from clean samples, we can correct the labels of noisy samples, further improving the model's performance against label noise. As a result, we propose a semi-supervised method.

Semi-supervised Method Based on RML. For the semi-supervised method based on RML, we separate noisy and clean samples into two groups. Existing selection methods usually adopt several hyperparameters and need to be adjusted according to the noise rate of different training datasets. To achieve an efficient and simple selection approach, we propose a new strategy, which uses f_θ and $g_{\theta'}$ to select samples. For a training sample $(\boldsymbol{x}, \tilde{y}) \in \tilde{\mathcal{D}}$, if its prediction label for both f_θ and $g_{\theta'}$ is equal to $\tilde{y}$, then it is reserved in $\tilde{\mathcal{D}}_{labeled}$. Otherwise, it is stored in $\tilde{\mathcal{D}}_{unlabeled}$.

The semi-supervised model is trained by two steps: common training and semi-supervised training. In common training, f_θ is trained with RML while $g_{\theta'}$ is updated by Eq. (5). In semi-supervised training, we use the MixMatch strategy to fully utilize the information of unlabeled samples and further improve the model performance. The detailed procedure can be found in Alg. 2 of Appendix .

Experiments

This section presents the training settings in the first part and then shows the experimental comparison between the proposed method. The last part discusses the impact of some operations in RML on the model performance and analyzes the experimental parameter settings.

Experimental Setup

This section briefly describes some of the experimental settings, including the datasets, network structure, and baselines. Details can be found in Appendix .

Datasets. We perform experiments on synthetic datasets and real-world datasets. For experiments on synthetic datasets, we choose two commonly used datasets CIFAR-10 and CIFAR-100 with different rates of symmetric label noise, pairflip label noise, and instance-dependent label noise (Xia et al. 2020). The generation of label noise in

Dataset	CIFAR-10			CIFAR-100		
Method / Symmetric	0.2	0.5	0.8	0.2	0.5	0.8
CE	86.5 ± 0.6	80.6 ± 0.2	63.7 ± 0.8	57.9 ± 0.4	47.3 ± 0.2	22.3 ± 1.2
MixUp	93.2 ± 0.3	88.2 ± 0.3	73.3 ± 0.3	69.5 ± 0.2	57.1 ± 0.6	34.1 ± 0.6
DivideMix	95.6 ± 0.1	94.6 ± 0.1	92.9 ± 0.3	75.3 ± 0.1	72.7 ± 0.6	56.4 ± 0.3
ELR+	94.9 ± 0.2	93.6 ± 0.1	90.4 ± 0.2	75.5 ± 0.2	71.0 ± 0.2	50.4 ± 0.8
PES	95.9 ± 0.1	95.1 ± 0.2	93.1 ± 0.2	77.4 ± 0.3	74.3 ± 0.6	61.6 ± 0.6
RML-Semi (Ours)	$\mathbf{96.5 \pm 0.2}$	$\mathbf{95.7 \pm 0.5}$	$\mathbf{93.9 \pm 0.2}$	$\mathbf{78.9 \pm 0.3}$	$\mathbf{77.8 \pm 0.7}$	$\mathbf{64.1 \pm 0.3}$

Table 2: Average test accuracy (%) comparison with state-of-the-art methods with semi-supervised strategy on CIFAR-10 and CIFAR-100. The mean and standard deviation over 5 runs are presented. Baselines are from (Bai et al. 2021).

Dataset	CIFAR-10			CIFAR-100		
Method / Noise	Instance-0.2	Instance-0.4	Pairflip-0.45	Instance-0.2	Instance-0.4	Pairflip-0.45
CE	87.5 ± 0.5	78.9 ± 0.7	74.9 ± 0.7	56.8 ± 0.4	48.2 ± 0.5	38.5 ± 0.6
MixUp	93.3 ± 0.2	87.6 ± 0.5	82.4 ± 1.0	67.1 ± 0.1	55.0 ± 0.1	44.2 ± 0.5
DivideMix	95.5 ± 0.1	94.5 ± 0.2	85.6 ± 1.7	75.2 ± 0.2	70.9 ± 0.1	48.2 ± 1.0
ELR+	94.9 ± 0.1	94.3 ± 0.2	86.1 ± 1.2	75.8 ± 0.1	74.3 ± 0.3	65.3 ± 1.3
PES	95.9 ± 0.1	95.3 ± 0.1	94.5 ± 0.3	77.6 ± 0.3	76.1 ± 0.4	73.6 ± 1.7
RML-Semi (Ours)	$\mathbf{96.3 \pm 0.2}$	$\mathbf{95.5 \pm 0.3}$	$\mathbf{95.2 \pm 0.3}$	$\mathbf{78.4 \pm 0.2}$	$\mathbf{76.8 \pm 0.5}$	$\mathbf{75.9 \pm 0.3}$

Table 3: Average test accuracy (%) comparison with state-of-the-art methods with semi-supervised strategy on CIFAR-10 and CIFAR-100. The mean and standard deviation over 5 runs are presented. Baselines are from (Bai et al. 2021).

our experiments follows (Bai et al. 2021). For the real-world datasets, we choose Clothing1M and WebVision.

Network Structure. All our experiments are performed on Ubuntu 20.04.3 LTS workstations with Intel Xeon 5120 and 5×3090 by PyTorch. To compare with baselines in common experimental results of the model without the semi-supervised strategy, we select ResNet-18 for the CIFAR-10 dataset and ResNet-34 for the CIFAR-100 dataset based on (Bai et al. 2021). For semi-supervised experiments, ResNet-18 is used as the backbone for both two datasets. For the real-world datasets, experiments on Clothing1M adopt a pre-trained ResNet-50, while the experiments on WebVision use ResNet-50 trained from scratch.

Baselines. We perform two groups of experiments on synthetic and real-world datasets. One trains the model directly, while the other adopts a semi-supervised strategy. For a fair comparison, we compare with different baselines with the same experimental settings. For approaches without semi-supervised strategy, we choose Forward (Patrini et al. 2017), Co-teaching (Han et al. 2018), Joint-Optim (Tanaka et al. 2018), MLNT (Li et al. 2019), T-revision (Xia et al. 2019), PCIL (Yi and Wu 2019), DMI (Xu et al. 2019), and PES (Bai et al. 2021). For semi-supervised group, we choose M-correction (Arazo et al. 2019), DivideMix (Li, Socher, and Hoi 2020), ELR+ (Liu et al. 2020), Sel-CL (Li et al. 2022), NCR (Iscen et al. 2022), and UNICON (Karim et al. 2022).

Results Comparison on Synthetic Datasets

We perform two groups of experiments on synthetic datasets. Tab. 1 shows the experimental comparisons on CIFAR-10 and CIFAR-100 without the semi-supervised strategy. For the experiments on CIFAR-10, we set k to 60 for a symmetric label noise ratio of 0.8. For instance-dependent and symmetric label noise with 0.2 ratio, the k is 600. The remaining experiments on CIFAR-10 adopt $k = 200$. The experiments on CIFAR-100 use $k = 50$ when the noise rate is 0.2. For the experiments on CIFAR-100 with a noise rate of 0.8, k is 6. For the rest of the experiments, k is set to 20. Compared to existing methods, RML helps the model achieve better performance on the two datasets with different noise rates and noise types. RML increases the average test accuracy by about 1% on CIFAR-10 and by about 6% on CIFAR-100. The improvement of the model trained by RML is remarkable when datasets have a higher noise rate, especially on CIFAR-100. Compared with the traditional training model, as shown in Tab. 2 and Tab. 3, the semi-supervised strategy further improves the model performance against label noise and achieves about 1% average test accuracy improvement on the two datasets, verifying the effectiveness of our proposed RML on synthetic datasets.

Method w/o Semi	Accuracy	Method w/ Semi	Accuracy
CE	69.21	DivideMix	74.76
Joint-Optim	72.16	ELR+	74.81
DMI	72.46	PES	74.99
MLNT	73.47	UNICON	74.98
PCIL	73.49	NCR	74.6
RML (Ours)	**73.54**	RML-Semi (Ours)	**75.14**

Table 4: Average test accuracy (%) comparison on Clothing1M. Baselines are from (Karim et al. 2022).

Results Comparison on Real-world Datasets

For real-world datasets, it is difficult to obtain noise rates and types. Therefore, methods should be able to handle various types of label noises and uncertain noise rates. To fully evaluate the effect of RML, we perform experiments on two real-world datasets and compare results with baselines.

Method w/o Semi	WebVision	ILSVRC12	Method w/ Semi	WebVision	ILSVRC12
Forward	61.12	57.36	DivideMix	77.32	75.20
Meta-Net	63.00	57.80	ELR+	77.78	70.29
Iterative-CV	65.24	61.60	UNICON	77.60	75.29
Co-training	63.58	61.48	Sel-CL	79.96	76.84
RML (Ours)	**73.32**	**75.04**	RML-Semi(Ours)	**81.34**	**77.38**

Table 5: Average test accuracy (%) comparison on WebVision. Baselines are from (Bai et al. 2021; Karim et al. 2022).

Dataset	CIFAR-10			CIFAR-100		
Method / Noise	Instance-0.4	Symmetric-0.5	Pairflip-0.45	Instance-0.4	Symmetric-0.5	Pairflip-0.45
w/o Loss Processing	88.47 ± 0.37	86.73 ± 0.33	87.75 ± 0.71	65.43 ± 0.45	62.31 ± 0.87	56.36 ± 1.23
w/o Median Operation	85.32 ± 0.54	83.61 ± 0.43	85.77 ± 0.25	64.37 ± 0.56	61.81 ± 0.59	55.32 ± 1.17
RML (Ours)	$\mathbf{90.73 \pm 0.44}$	$\mathbf{89.13 \pm 0.21}$	$\mathbf{90.14 \pm 0.34}$	$\mathbf{71.12 \pm 1.12}$	$\mathbf{67.05 \pm 0.31}$	$\mathbf{64.67 \pm 0.54}$

Table 6: Ablation study with different training settings. The mean and standard deviation over 5 runs are presented.

Performance on Clothing1M. Tab. 4 compares the model performance with baselines on Clothing1M. We set k to 200 for each 1000 selected training batches. Compared to existing traditionally trained methods, the model trained with RML achieves better performance. For the semi-supervised experiments on Clothing1M, the semi-supervised model with RML further strengthens the model performance. Tab. 4 verifies that the semi-supervised model with RML is able to deal with label noise in such a complex real-world Clothing1M dataset. Compared to the current semi-supervised methods, the semi-supervised model with RML increases the test accuracy by 0.15%.

Performance on WebVision. The WebVision dataset is more challenging because we need to train models from scratch. Due to the unbalanced samples of each class, we select 75% losses of training samples and divide them into 6 groups to train the model. For the traditional training methods in Tab. 5, the improvement of the model with RML is remarkable compared to existing methods. Our method achieves an improvement in accuracy of 8% and 14% on WebVision and ILSVRC12, respectively. The comparison between Co-training and RML shows that RML addresses the issues of the existing small-loss criterion and is better at combating label noise in real-world datasets. Tab. 5 demonstrates that the semi-supervised method with RML can protect the model from label noise in real-world datasets and improve the model generalization. The semi-supervised model achieves about 1.5% improvement in test accuracy on WebVision and 0.54% improvement on ILSVRC12, obviously better than the current state-of-the-art methods.

Ablation Studies

In this part, we discuss the impact of loss processing, median operation and settings of n on experimental results.

Loss Processing. As shown in Tab. 6, we compare the experimental results of the model for RML with and without loss processing on CIFAR-10 and CIFAR-100 with various types of label noises, indicating that the proposed loss processing apparently improves the model's performance.

Median Operation. The median operation plays an essential role in RML to combat label noise. To verify the function of the median operation, we perform experiments on CIFAR-10 and CIFAR-100. For groups of experiments without the median operation, we simply compute the mean of the selected sample losses. Tab. 6 shows the experimental results of RML in two cases. Compared with the mean loss, RML with the regroup median operation helps the model achieve better performance, which indicates that the median operation can mitigate the negative impact of selected noisy samples and provide a more appropriate loss estimation for noisy samples than the mean loss.

Parameter Settings of n. In RML, n is the window size of the median operation and should be an even number. To explore the appropriate setting of n, we perform experiments on CIFAR-10 with various types of label noises to evaluate the impact of different values of n on the model convergence speed and performance. As shown in Fig. 3 and Tab. 8 in Appendix, when the value of n is higher, the convergence speed is slower because the higher n is, the smaller the estimated loss is. To take into account the model performance and the convergence speed, we choose $n = 6$ on all datasets.

Conclusions

In this work, we propose RML, a novel method to combat label noise. We analyze the shortcomings of existing methods and introduce the motivation of our proposed method. Then, we present the details of RML and explain some of its main operations. Based on RML, we propose a semi-supervised method based on RML to further improve the model performance against label noise. To verify the effectiveness of RML, we perform a large number of experiments on synthetic datasets with various types of label noises and real-world datasets with unknown noise types and noise rates. These experiments show that RML corrects the distorted losses of noisy samples and provides an appropriate estimation for each training sample. Compared to state-of-the-art methods, these results show that RML further improves the model performance against label noise.

Acknowledgments

This work was supported in part by Macau Science and Technology Development Fund under SKLIOTSC-2021-2023, 0072/2020/AMJ and 0022/2022/A1; in part by Research Committee at University of Macau under MYRG2022-00152-FST and MYRG-GRG2023-00058-FST-UMDF; in part by Natural Science Foundation of China under 61971476; and in part by Alibaba Group through Alibaba Innovative Research Program.

References

Arazo, E.; Ortego, D.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2019. Unsupervised Label Noise Modeling and Loss Correction. In *Proceedings of ICML, June 9-15, 2019, Long Beach, CAL, USA*, 312–321.

Arpit, D.; Jastrzundefinedbski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; and Lacoste-Julien, S. 2017. A Closer Look at Memorization in Deep Networks. In *Proceedings of ICML - Volume 70*, 233–242. JMLR.org.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. In *Proceedings of NeuIPS, December 6-14, 2021, virtual*, 24392–24403.

Berthelot, D.; Carlini, N.; Goodfellow, I. J.; Papernot, N.; Oliver, A.; and Raffel, C. 2019. MixMatch: A Holistic Approach to Semi-Supervised Learning. In *Proceedings of NeuIPS, December 8-14, 2019, Vancouver, BC, Canada*, 5050–5060.

Catoni, O. 2012. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, 1148–1185.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *Proceedings of ICLR, May 3-7, 2021, Virtual Event, Austria*.

Cui, X.; Saon, G.; Nagano, T.; Suzuki, M.; Fukuda, T.; Kingsbury, B.; and Kurata, G. 2022. Improving Generalization of Deep Neural Network Acoustic Models with Length Perturbation and N-best Based Label Smoothing. *arXiv preprint arXiv:2203.15176*.

Frénay, B.; and Verleysen, M. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25: 845–869.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the CVPR, June 7-12, 2015, Boston, MA, USA*, 1440–1448.

Gui, X.; Wang, W.; and Tian, Z. 2021. Towards Understanding Deep Learning from Noisy Labels with Small-Loss Criterion. In *Proceedings of IJCAI, August 19-27, 2021, Virtual Event, Montreal, Canada*, 2469–2475.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of NeuIPS, December 3-8, 2018, Montréal, Canada*, 8536–8546.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the ICCV, October 22-29, 2017, Venice, Italy*, 2961–2969.

Iscen, A.; Valmadre, J.; Arnab, A.; and Schmid, C. 2022. Learning with Neighbor Consistency for Noisy Labels. In *Proceedings of CVPR, June 18-24, 2022, New Orleans, LA, USA*, 4662–4671.

Jenni, S.; and Favaro, P. 2018. Deep Bilevel Learning. In *Proceedings of CVPR, September 8-14, 2018, Munich, Germany*, 632–648.

Karim, N.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning. In *Proceedings of the CVPR, June 18-24, 2022, New Orleans, LA, USA*, 9676–9686.

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

Lecué, G.; and Lerasle, M. 2020. Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics*, 48(2): 906 – 931.

Li, J.; Socher, R.; and Hoi, S. C. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of ICLR, April 26-30, 2020, Addis Ababa, Ethiopia*, 824–832.

Li, J.; Wong, Y.; Zhao, Q.; and Kankanhalli, M. S. 2019. Learning to Learn From Noisy Labeled Data. In *Proceedings of CVPR, June 16-20, 2019, Long Beach, CA, USA*, 5051–5059.

Li, S.; Xia, X.; Ge, S.; and Liu, T. 2022. Selective-Supervised Contrastive Learning with Noisy Labels. In *Proceedings of CVPR, June 18-24, 2022, New Orleans, LA, USA*, 316–325.

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*.

Li, X.; Liu, T.; Han, B.; Niu, G.; and Sugiyama, M. 2021. Provably end-to-end label-noise learning without anchor points. In *Proceedings of ICML, July 18-24, 2021, Virtual Event, USA*, 6403–6413.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. In *Proceedings of NeuIPS, December 6-12, 2020, virtual, USA*, 20331–20342.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? In *Proceedings of NeuIPS, December 8-14, 2019, Vancouver, BC, Canada*, 4696–4705.

Nemirovskij, A. S.; and Yudin, D. B. 1985. Problem complexity and method efficiency in optimization. *Siam Review*, 27(2): 264–266.

Ortego, D.; Arazo, E.; Albert, P.; O'Connor, N. E.; and McGuinness, K. 2021. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the CVPR, June 19-25, 2021, virtual, USA*, 6606–6615.

Patrini, G.; Rozza, A.; Menon, A. K.; Nock, R.; and Qu, L. 2017. Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. In *Proceedings of CVPR, July 21-26, 2017, Honolulu, HI, USA*, 2233–2241.

Rasmussen, C. 1999. The infinite Gaussian mixture model. In *Proceedings of NeuIPS, November 29 - December 4, 1999, Denver, Colorado, USA*, volume 12, 554–560.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of ICCV, June 27-30, 2016, Las Vegas, NV, USA*, 779–788.

Song, H.; Kim, M.; Park, D.; and Lee, J. 2020. Learning from Noisy Labels with Deep Neural Networks: A Survey. *CoRR*, abs/2007.08199.

Staerman, G.; Laforgue, P.; Mozharovskyi, P.; and d'Alché Buc, F. 2021. When ot meets mom: Robust estimation of Wasserstein distance. In *Proceedings of AISTATS, April 13 - 15, 2021, Virtual Conference, USA*, 136–144.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of CVPR, June 7-12, 2015, Boston, MA, USA*, 1–9.

Tanaka, D.; Ikami, D.; Yamasaki, T.; and Aizawa, K. 2018. Joint Optimization Framework for Learning With Noisy Labels. In *Proceedings of CVPR, June 18-22, 2018, Salt Lake City, UT, USA*, 5552–5560.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of NeuIPS, December 4-9, 2017, Long Beach, CA, USA*, 1195–1204.

Van Rooyen, B.; Menon, A.; and Williamson, R. C. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Proceedings of NeuIPS, December 7-12, 2015, Montreal, Quebec, Canada*, 10–18.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2021. Sample selection with uncertainty of losses for learning with noisy labels. *arXiv preprint arXiv:2106.00445*.

Xia, X.; Liu, T.; Han, B.; Wang, N.; Gong, M.; Liu, H.; Niu, G.; Tao, D.; and Sugiyama, M. 2020. Part-dependent label noise: Towards instance-dependent label noise. In *Proceedings of NeuIPS, December 6-12, 2020, Virtual Event, USA*, 7597–7610.

Xia, X.; Liu, T.; Wang, N.; Han, B.; Gong, C.; Niu, G.; and Sugiyama, M. 2019. Are anchor points really indispensable in label-noise learning? In *Proceedings of NeuIPS, December 8-14, 2019, Vancouver, BC, Canada*, 6835–6846.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification. In *Proceedings of CVPR, June 7-12, 2015, Boston, MA, USA*, 2691–2699.

Xu, Y.; Cao, P.; Kong, Y.; and Wang, Y. 2019. L_DMI: A Novel Information-theoretic Loss Function for Training Deep Nets Robust to Label Noise. In *Proceedings of NeuIPS, December 8-14, 2019, Vancouver, BC, Canada*, 6222–6233.

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 *Proceedings of CVPR, June 19-25, 2021, virtual, USA*, 5192–5201.

Yi, K.; and Wu, J. 2019. Probabilistic End-To-End Noise Correction for Learning With Noisy Labels. In *Proceedings of CVPR, June 16-20, 2019, Long Beach, CA, USA*, 7017–7025.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *Proceedings of ICML, June 9-15, 2019, Long Beach, California, USA*, 7164–7173.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *Proceedings of ICLR*.

Zhang, C.-B.; Jiang, P.-T.; Hou, Q.; Wei, Y.; Han, Q.; Li, Z.; and Cheng, M.-M. 2021. Delving deep into label smoothing. *IEEE Transactions on Image Processing*, 30: 5984–5996.

Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018. Context encoding for semantic segmentation. In *Proceedings of CVPR, June 18-22, 2018, Salt Lake City, UT, USA*, 7151–7160.

Zhang, X.; Zou, J.; He, K.; and Sun, J. 2015. Accelerating very deep convolutional networks for classification and detection. *IEEE transactions on pattern analysis and machine intelligence*, 38(10): 1943–1955.

Zhang, Y.; Niu, G.; and Sugiyama, M. 2021. Learning noise transition matrix from only noisy labels via total variation regularization. In *Proceedings of ICML, July 18-24, 2021, Virtual Event, USA*, 12501–12512.

Zhu, Z.; Liu, T.; and Liu, Y. 2021. A second-order approach to learning with instance-dependent label noise. In *Proceedings of CVPR, June 19-25, 2021, virtual Event, USA*, 10113–10123.

Zubiaga, A.; Liakata, M.; Procter, R.; Bontcheva, K.; and Tolmie, P. 2015. Crowdsourcing the annotation of rumourous conversations in social media. In *Proceedings of the WWW, May 18-22, 2015, Florence, Italy*, 347–353.