LEARNING WITH NON-UNIFORM LABEL NOISE: A CLUSTER-DEPENDENT SEMI-SUPERVISED APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning with noisy labels is a challenging task in machine learning. Most existing methods explicitly or implicitly assume uniform label noise across all samples. In reality, label noise can be highly non-uniform in the feature space, e.g. with higher error rate for more difficult samples. Some recent works consider instance-dependent label noise but they require additional information such as some cleanly labeled data and confidence scores, which are usually unavailable or costly to obtain. In this paper, we consider learning with non-uniform label noise that requires no such additional information. We propose a cluster-dependent sample selection algorithm followed by a semi-supervised training mechanism based on the cluster-dependent label noise. Inspired by stratified sampling, the proposed sample selection method increases the consistency of sample space by forcing the selection of clean samples from the entire feature space. Despite its simplicity, the proposed method can distinguish clean data from the corrupt ones more precisely and achieve state-of-the-art performance on most image classification benchmarks, especially when the number of training samples is small and the noise rate is high.

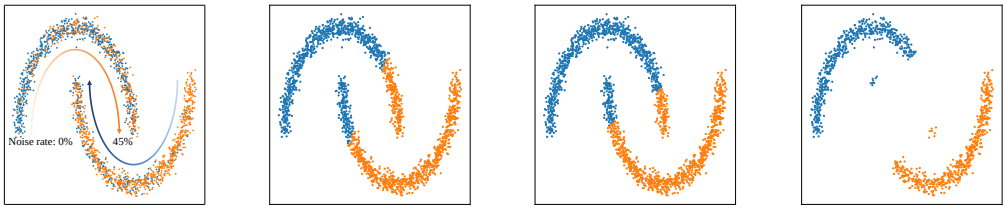
1 INTRODUCTION

Deep Neural Networks (DNNs) have achieved great success in various machine learning tasks, such as in computer vision, natural language processing, and information retrieval. Unfortunately, their successes heavily rely on the carefully labeled data, which are expensive and time-consuming to collect. Online queries (Xie et al., 2019) and crowdsourcing (Yu et al., 2018) are cheap alternatives, which would produce datasets with noisy labels. Furthermore, the datasets in medical applications are typically small and domain expertise is required to annotate medical data, which however often suffers from inter- and intra-observer variability (Han et al., 2020; Xue et al., 2022). Song et al. (2022) reports that the ratio of corrupted labels in real-world datasets range from 8.0% to 38.5%.

Due to the universal approximation ability of DNNs (Hornik et al., 1989), they can easily memorize and eventually overfit to the corrupted labels, leading to poor generalization (Zhang et al., 2017). Efforts have been taken to robust learning paradigms under noisy labels (Frénay & Verleysen, 2014; Han et al., 2020; Song et al., 2022). Generally, existing methods on learning with noisy labels can be categorized into two groups: loss correction methods (Patrini et al., 2017; Xia et al., 2019; Yao et al., 2020; Xu et al., 2019; Xia et al., 2020b; Berthon et al., 2021) and sample selection methods (Han et al., 2018; Yu et al., 2019; Li et al., 2020; Bai et al., 2021).

Methods in the first category mainly model label noise with label transition matrix. Patrini et al. (2017); Xia et al. (2019) and Yao et al. (2020) assumed that label noises were class-dependent but instance-independent (note as class-dependent noise, CDN). Xia et al. (2020b) and Berthon et al. (2021) proposed to model instance-dependent noise (IDN). Obviously, the IDN transition matrices are more realistic (see Figure 2) but unidentifiable in general (Liu, 2022), which require a large number of parameters to be estimated. Additional information (Berthon et al., 2021) or extra assumption (Xia et al., 2020b) is required to estimate instance-dependent transition matrix.

The methods of second type are designed to select confident clean samples from noisy datasets based on the memorization effect of DNNs (Arpit et al., 2017), which tend to learn simple patterns first before fitting the corrupt samples. Han et al. (2018) and Yu et al. (2019) train two networks simultaneously where each network selects small-loss samples to train the other one. Furthermore,



(a) Moon Data (b) T-Revision (c) DivideMix (d) Clean data selected by DivideMix

Figure 1: An illustration of the moon dataset with gradually changed noise rate. (a) The moon dataset. The local noise rate varies from 0% to 45% gradually, with the average of 27.15%. (b) Visualization result of T-Revision (Xia et al., 2019) on the noisy moon data. T-Revision performs loss correction by the noise transition matrix estimated a slack variable. (c) Visualization result of DivideMix (Li et al., 2020) on the noisy moon data. DivideMix selects clean data through the small-loss trick and then train the network with the semi-supervised paradigm. (d) Clean data selected by DivideMix. The method fails to select samples from the heavy noise region, which shows the inconsistency of sample selection under non-uniform noise. Besides, Appendix A shows the effective experiment results based on the proposed cluster-dependent sample selection strategy.

semi-supervised technologies was used to explore both confident clean samples (as labeled data) and corrupt samples (as unlabeled data) (Li et al., 2020; Bai et al., 2021).

Existing works explicitly or implicitly rely on the assumption of uniform noise, while the noise rate of real dataset range from $25\% \sim 60\%$ in Figure 2. The methods of CDN transition matrix explicitly consider all the samples in the same class have same noise rate and methods of IDN need additional information and extra assumption, which are not realistic and have poor performance experimentally. Without explicit assumption of uniform noise rate, the small-loss trick would still select simple patterns first but regard all samples in the hard regions as corrupt data (though part of the samples have correct labels), resulting in inconsistent sample distribution (Cheng et al., 2020). Notably, Liu & Wang (2021) reports that increasing label noise to balance can return more accurate models.

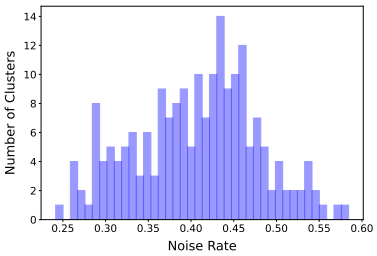


Figure 2: The noise rate distribution of 200 clusters on CIFAR-10 Noise dataset Wei et al. (2022). Pre-trained features are used to produces the clusters.

To empirically verify the poor performance of existing methods on non-uniform noisy data, we conduct a demo experiment on moon data with gradually changed noise rate, shown in Figure 1. Non-uniform label noise (IDN, noise rate range from 0% to 45%) is added to the moon dataset and different methods (Xia et al., 2019; Li et al., 2020) are employed to learn the noise-robust classifier. As illustrated in Figures 1b and 1c, both existing loss correction and sample selection methods fail to classify the moon dataset. Clearly, T-Revision with transition matrix explicitly suppose that all the samples share the same noise rate, which is contradictory to the real noise distribution and leads to the poor performance. Although the small-loss trick of DivideMix does not require the same noise rate, no sample in the high noise rate region is selected despite the existence of correct samples, resulting in poor decision boundary.

To address the above problem of inconsistent sample selection problem, we assume that label noise is dependent on the clusters of features (Cluster-Dependent Noise, CluDN, Definition 3.1 for more details), i.e. the samples share the same noise rate in the same cluster while have different noise rate among clusters. Based on the idea of stratified sampling, we propose a cluster-dependent sample selection algorithm followed by a semi-supervised training mechanism, which is named as ClusterMix. Cluster-dependent sample selection and semi-supervised training are conducted in turn and precise clean data would be selected progressively. Based on the cluster-dependent sample selection, samples in both simple and hard regions can be selected which eliminate the inconsistency defect of existing sample selection methods. Experimental results on real non-uniform datasets verify the

effectiveness of our proposed method, which outperforms all the baseline methods when the training set is small and noise rate is high.

The rest of the paper is organized as follows. In Section 2, we review the related work briefly. In Section 3, we first give the our intuition analysis of cluster-dependent label noise and then introduce our proposed ClusterMix algorithm in details. Section 4 displays the experimental results on various image classification benchmarks. Finally, concluding remarks are given in Section 5.

2 Related Work

Learning with label noise has been extensively studied (Frénay & Verleysen, 2014; Han et al., 2020; Song et al., 2022). Generally, existing methods on learning with noisy labels can be categorized into two groups: loss correction methods and sample selection methods. The loss correction methods aim to design risk-consistent training mechanism based on loss correction, while the sample selection methods try to select clean samples from the entire noisy dataset and recently combine with semi-supervised learning. As the experimental results demonstrates the effectiveness of sample selection methods, we will focus on the sample selection methods in this paper.

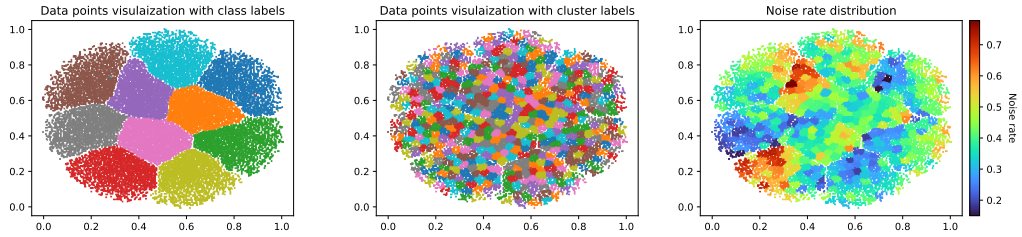
Learning with loss correction. To deal with label noise, methods have been proposed to model the relationship between clean labels and noisy labels by transition matrix and build loss functions to correct it. Patrini et al. (2017); Goldberger & Ben-Reuven (2017); Xia et al. (2019) and Zhu et al. (2021) provide methods to estimate transition matrix and use forward and backward procedures for loss correction. Besides, Xu et al. (2019) proposes a information-theoretic loss function, which is provably robust to instance-independent label noise. However, the instance-independent label noise used in these methods is inconsistent with the non-uniform label noise in reality. To deal with the instance-dependent label noise, methods (Xia et al., 2020b; Yang et al., 2021; Berthon et al., 2021) estimate a separate transition matrix for each samples. Nevertheless, the number of matrices parameters is large, resulting in large estimation error in practical applications. For example, given a dataset with 10000 samples and 10 classes, 10^6 parameters are needed. Furthermore, Xia et al. (2020a); Zhu et al. (2021) and Zhang et al. (2021) propose to estimate cluster-dependent transition matrices. Although these methods have made certain progress, they are hard to handle a large number of classes and struggling to estimate accurately for heavy noise.

Sample selection. The second strand tries to select clean samples from the noisy dataset, by exploiting the memorization effect of DNNs (Arpit et al., 2017). A common method is to treat samples with small loss as clean ones (Shen & Sanghavi, 2019). To utilize the property, Co-teaching (Han et al., 2018) train two networks simultaneously and let them select clean samples for each other within each mini-batch. Co-teaching+ (Yu et al., 2019) improves Co-teaching by maintaining disagreement between the two networks. MentorNet (Jiang et al., 2018) use a mentor network to select confident clean samples for the training of student network. To utilize the corrupt samples, some works (Li et al., 2020; Bai et al., 2021; Chen et al., 2021) treat un-selected data as unlabeled data and conduct semi-supervised learning (SSL) mechanism to train networks. DivideMix (Li et al., 2020) employs two Gaussian Mixture Model (GMM) to select clean samples and MixMatch (Berthelot et al., 2019) strategy to leverage corrupt examples with SSL frameworks. PES (Bai et al., 2021) select confident clean samples by progressive early stopping. SOP (Liu et al., 2022) propose to model the label noise and learn to separate it from the data by over-parameterization. However, existing methods usually adopt unified selection criteria across all samples, where all the hard samples would be regarded as corrupt ones, resulting in inconsistent data space. In contrast, our method is designed to select samples from each cluster separately. With variable selection criteria on different clusters, our method is able to better exploit the whole data space and thus achieve superior performance.

3 Method

3.1 Preliminaries

Suppose $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ are drawn from an unknown joint distribution $P_{\mathcal{D}}$, where $\mathcal{X} \subset \mathbb{R}^d$ is the feature space and $\mathcal{Y} \subset \{0, 1\}^C$ is the clean label space in a one-hot manner, where d is the feature dimension and C is the number of categories. Then $\mathcal{D} = \{(x_n, y_n)\}_{n\in[N]}$ is a clean training dataset with N samples.



(a) Data points with class index. (b) Data points with cluster index. (c) Noise rate distribution.

Figure 3: Visualization of CIFAR-10 data points and the noise distribution of CIFAR-10N Worst Label. The data points are obtained by performing T-SNE dimension reduction algorithm on 512-d features extracted by Resnet 18. The noise rate for each point is calculated by the average corrupt rate among its cluster.

Instead of observing clean label Y, people can only collect noisy label $\tilde{Y}$ in various real-world scenarios. Let noisy dataset $\tilde{\mathcal{D}} = \{(x_n, \tilde{y}_n)\}_{n\in[N]}$. We denote $\tilde{y}_n$ as corrupt label if $\tilde{y}_n \neq y_n$ and clean otherwise. The noise rate for sample x is denote as $\tau(x) = \Pr(\tilde{y} \neq y|x)$. And the noise rate for the whole dataset is,

$$\tau(\tilde{\mathcal{D}}) = \frac{|\{(x_n, \tilde{y}_n) : (x_n, \tilde{y}_n) \in \tilde{\mathcal{D}} \ \ \& \ \ \tilde{y}_n \neq y_n\}|}{|\tilde{\mathcal{D}}|}$$

In the sample selection methods, the entire dataset $\tilde{\mathcal{D}}$ would be grouped to K clusters and finally divided into a clean subset $\tilde{\mathcal{D}}_{\text{clean}} = \{(x_n, \tilde{y}_n) : (x_n, \tilde{y}_n) \in \tilde{\mathcal{D}} \ \ \& \ \ \tilde{y}_n = y_n\}$ and a corrupted subset $\tilde{\mathcal{D}}_{\text{corrupt}} = \{(x_n, \tilde{y}_n) : (x_n, \tilde{y}_n) \in \tilde{\mathcal{D}} \ \ \& \ \ \tilde{y}_n \neq y_n\}$. Our aim is to learn a robust classifier $f(\cdot, \theta) : X \to Y$ based only on corrupted dataset $\tilde{\mathcal{D}}$.

3.2 Intuitions

As shown in Figure 2, the CDN model, which assumes that all samples in the same class share same noise rate, is unrealistic in general scenarios. Besides, IDN model is proposed to fit the real noise but suffers the non-identifiability (Liu, 2022). Specifically, different combination of clean posterior $P(Y|X)$ and instance noise rate $\tau(X)$ can lead to the same noisy posterior $P(\tilde{Y}|X)$. To make a trade-off between flexibility and identifiability, we assume that the real label noise is dependent on the clusters. The samples in the same cluster share the same noise rate, as stated in Definition 3.1.

Definition 3.1 (Cluster-dependent Label Noise Model)**.** Suppose a noisy dataset $\tilde{\mathcal{D}}$ can be divided into multiple clusters based on features. **Cluster-Dependent Noise Model (CluDN)** assumes that the samples in each cluster share the same corrupt probability and the corrupt probabilities among clusters can vary greatly.

In the paper, clusters are treated as the finer hierarchy than the classes (illustrated in Figures 3a and 3b), i.e. one class can contain multiple clusters. On the contrary, clusters are not always the sub-group of the classes, as the clusters on the class boundary can cross two- or multi- classes. To further explore the validity of the proposed noise model, corresponding noise rate distribution are visualized in Figure 3c, which illustrates the non-uniform noise distribution within each class. Besides, Appendix E displays several images of two example clusters, which gives a potential source of non-uniform noise on real dataset.

To design a robust learning mechanism under non-uniform noise, we combine the CluDN model with existing sample selection methods. In CluDN model, samples among different clusters have different patterns and thus unified selection criteria among clusters is inappropriate under non-uniform noise. Based on the idea of stratified sampling, we propose to select clean samples for each cluster separately and aggregate them together to get the clean data $\tilde{\mathcal{D}}_{\text{clean}}$ and corrupt data $\tilde{\mathcal{D}}_{\text{corrupt}}$. More details will be discussed in the following subsection.

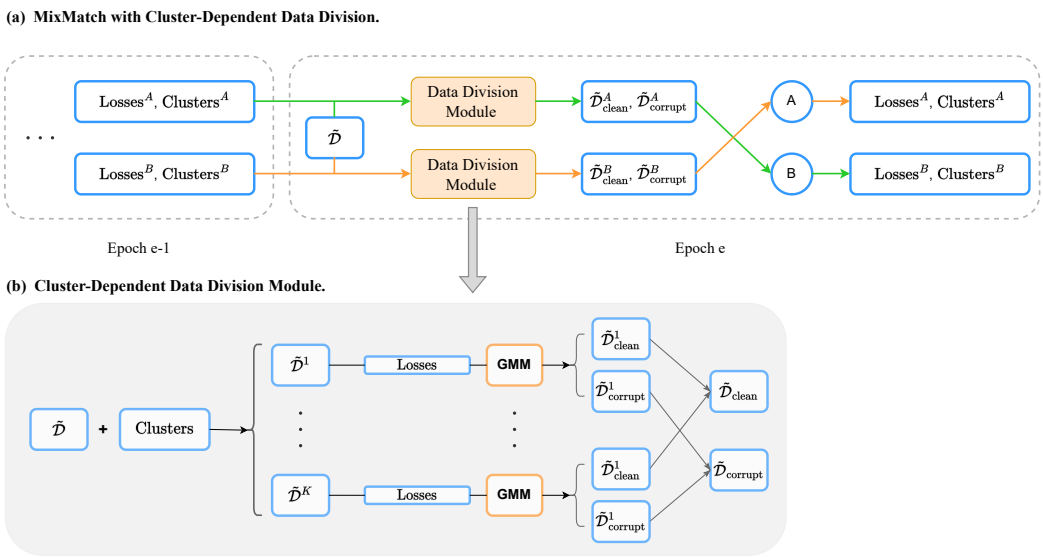


Figure 4: An overview of proposed ClusterMix, which is robust to the non-uniform noise. (a) is the semi-supervised co-training paradigm with cluster-dependent data division modules. (b) presents the details of the cluster-dependent data division module. Using the small-loss trick, A Gaussian Mixture Model (GMM) will divide the data into clean and corrupt parts for each cluster separately. Afterwards, the clean and corrupt data in all clusters will merge respectively.

3.3 ClusterMix: Semi-supervised Learning with Cluster-Dependent Sample Selection

Based on CluLN, we propose ClusterMix, a robust semi-supervised algorithm with cluster-dependent sample selection module. An overview of the framework is shown in Figure 4. To avoid the accumulation of confirmation bias in self-training, we train two identical network with different random initialization simultaneously, following Han et al. (2018); Yu et al. (2019) and Li et al. (2020).

As shown in Figure 4, the algorithm consists of two parts, i.e. cluster-based data division module and SSL training paradigm, specifically MixMatch (Berthelot et al., 2019). The cluster-dependent data division module will first divide the entire dataset $\tilde{\mathcal{D}}$ into clean subset $\tilde{\mathcal{D}}_{\text{clean}}$ and corrupt subset $\tilde{\mathcal{D}}_{\text{corrupt}}$ based on cluster and loss information. Then in the second part, the clean set will be treated as labeled data while corrupt one as unlabeled data. The SSL module use the label co-refinement and co-guess techniques (Li et al., 2020) to generate pseudo labels and MixMatch technique (Berthelot et al., 2019) to perform SSL regularization.

Cluster-Dependent Data Division. As shown above, the samples in different clusters have different noise rate. And samples with high noise rate are more difficult to classify, i.e. all of them would have larger losses. Thus, the samples in the ambiguous feature space would all be regarded as corrupted instances, resulting in inconsistent data distribution. To address the challenge of non-uniform label noise, we propose to select clean and corrupt samples within clusters.

Inspired by the idea of divide-and-conquer, we propose to first divide the entire dataset into several cluster-based subset. According to the CluDN model, the samples in the same cluster have the same noise rate and thus have comparable prediction losses. Therefore by using the small-loss trick on each cluster, an independent 2-components GMM is used to fit the per-sample losses and the samples clean probability $\omega_i = p(g|l_i)$, where g is the component with lower mean and l_i is the loss of sample i. Finally, the clean samples ($\omega_i > 0.5$) in all clusters will merge together as the labeled data while the corrupted ones as the unlabeled data.

The visualization results in Appendix C illustrate the effectiveness of the proposed cluster-dependent sample selection strategy, which actually adds another dimension and thus have variable criteria for different clusters. The proposed method can select more precise and adequate clean samples and thus achieve better evaluation accuracy.

Algorithm 1: Cluster-Dependent Data Division Module.

Input: per-sample losses L, data clusters Clusters, Noisy dataset $\tilde{\mathcal{D}} = \{(x_n, \tilde{y}_n)\}_{n\in[N]}$, Clusters number K.

/* Divide the entire dataset and losses into clusters. */

$\tilde{\mathcal{D}}^1, \dots \tilde{\mathcal{D}}^K = \text{DataCluster}(\tilde{\mathcal{D}}, \text{Clusters})$;

$\tilde{L}^1, \dots \tilde{L}^K = \text{LossesCluster}(\text{L}, \text{Clusters})$;

for $k = 0;\ k < K;\ k+ = 1$ **do**

  /* Clean and corrupt samples selection for each cluster. */

  $\tilde{\mathcal{D}}^k_{clean}, \tilde{\mathcal{D}}^k_{corrupt} = \text{GMM}(\tilde{\mathcal{D}}^k, \text{L}^k)$;

end

/* Merge selected data together. */

$\tilde{\mathcal{D}}_{clean} = \text{Merge}(\tilde{\mathcal{D}}^1_{clean}, \dots, \tilde{\mathcal{D}}^K_{clean})$;

$\tilde{\mathcal{D}}_{corrupt} = \text{Merge}(\tilde{\mathcal{D}}^1_{corrupt}, \dots, \tilde{\mathcal{D}}^K_{corrupt})$;

return $\tilde{\mathcal{D}}_{clean}, \tilde{\mathcal{D}}_{corrupt}$.

Algorithm 2: ClusterMix: SSL with Cluster-Dependent Sample Selection

Input: Classifier $f_1(\cdot)$ with parameters Θ_1, $f_2(\cdot)$ with parameters Θ_2; feature extractors $g_1(\cdot)$, $g_2(\cdot)$; Cluster function $h(\cdot)$, Noisy dataset $\tilde{\mathcal{D}} = \{(x_n, \tilde{y}_n)\}_{n\in[N]}$, Clusters number K, Training epochs E.

1 WarmUp($\tilde{\mathcal{D}}$, Θ_1, Θ_2);

2 **for** $e <$ E **do**

3 $\text{L}_1, \text{Clusters}_1 = \text{CrossEntropyLoss}(f_2(X), Y), h(g_2(X), K)$;

4 $\text{L}_2, \text{Clusters}_2 = \text{CrossEntropyLoss}(f_1(X), Y), h(g_1(X), K)$;

5 **for** $i = 1, 2$ **do**

  /* Cluster-dependent data division module. */

6 $\tilde{\mathcal{D}}^{(i)}_{clean}, \tilde{\mathcal{D}}^{(i)}_{corrupt} = \text{DataDivision}(\text{L}_i, \text{Clusters}_i, \tilde{\mathcal{D}}, K)$;

  /* Label co-refinement and co-guessing. */

7 $\mathcal{X} \leftarrow \tilde{\mathcal{D}}^{(i)}_{clean}, \mathcal{U} \leftarrow \tilde{\mathcal{D}}^{(i)}_{corrupt}$;

8 $\mathcal{X}', \mathcal{U}' \leftarrow \mathcal{X}, \mathcal{U}$ // Samples MixMatch in mini-batch.

  /* Co-training. ($\sim i = 2$ if $i = 1$, vice versa.) */

9 Training parameters $\Theta_{\sim i}$ with CE loss on $\mathcal{X}'$ and MSE loss on $\mathcal{U}'$.

10 **end**

11 **end**

Pseudo label generation. To account for the label noise, label co-refinement is employed to modified the labeled data and label co-guessing is used to generate pseudo label for unlabeled data using the aggregate output of two networks. More details are shown in Li et al. (2020) and the step is used in Algorithm 2, line 7. By the label co-refinement and co-guessing above, pseudo labels are generated for both labeled dataset $\mathcal{X}$ and unlabeled dataset $\mathcal{U}$.

Semi-supervised learning - MixMatch. With the generated pseudo labels, MixMatch (Berthelot et al., 2019) is conducted for SSL, which utilizes unlabeled data by combining consistency regularization and entropy minimization with the MixUp (Zhang et al., 2018) augmentation. For a pair of samples $(x_1, \hat{y}_1)$ and $(x_2, \hat{y}_2)$, the mixed sample (x', y') is computed by

$$\lambda \sim \text{Beta}(\alpha, \alpha), \qquad \lambda' = \max(\lambda, 1-\lambda),$$
$$x' = \lambda' x_1 + (1-\lambda')\, x_2, \qquad y' = \lambda' \hat{y}_1 + (1-\lambda')\, \hat{y}_2.$$

The mixed sets are denoted as $\mathcal{X}'$ and $\mathcal{U}'$. Finally parameters Θ_1, Θ_2 are trained with cross-entropy (CE) loss on $\mathcal{X}'$ and mean-squared-error (MSE) loss on $\mathcal{U}'$.

Clustering. Different from the moon data shown in Figure 1 which can be grouped on the raw 2-dimensional data space, we use the features extracted by networks to divide the general dataset into clusters. Clustering and network training take turns in each epoch, where clusters are used for robust feature extraction and conversely extracted features would benefit the precise clustering. Various clustering methods have been proposed recently, e.g. partition-based clustering, spectral clustering, manifold clustering, and subspace clustering etc. As our method is robust to the clustering

Table 1: Test accuracy (%) with realistic label noise on CIFAR-N. For CIFAR-10N, we use noisy label aggregate ($\tau = 9.03\%$), random 1 ($\tau = 17.23\%$), and Worst ($\tau = 40.21\%$). And for CIFAR-100N, we use the fine noisy label with $\tau = 40.20\%$. **Bold** means the highest reported accuracy and underline is the second highest accuracy.

Methods	CIFAR-10N			CIFAR-100N
	Aggregate (9.03%)	Random 1 (17.23%)	Worst (40.21%)	Fine (40.20%)
CE (Standard)	89.87	84.15	76.86	55.96
T-Revision	89.39	87.99	82.10	54.45
PTD	89.93	89.83	80.16	16.01
ELR+	94.81	94.54	90.89	67.04
DivideMix	95.15	95.12	92.71	71.13
SOP	95.61	95.28	93.24	67.81
ClusterMix(ours)	**95.63**	**95.46**	**93.47**	**71.60**

algorithm, we adopt K-Means for convenience. Besides, experimental result shows the robustness of our method to the clusters number.

4 Experiments

4.1 Datasets and Implementation Details

Datasets: We evaluate our method on three real noisy dataset CIFAR-10N, CIFAR-100N (Wei et al., 2022), Clothing1M (Xiao et al., 2015), and (mini) WebVision (Li et al., 2017). CIFAR-10N and CIFAR-100N equipe the training sets of CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) with human-annotated real-world noisy labels coleected from Amazon Mechanical Turk. Both CIFAR-10 and CIFAR-100 contain 50K training images and 10K test images with size 32×32. CIFAR-10N provides several set of real-world noisy labels for training set, i.e. Aggregate (noise rate $\tau = 9.03\%$), Random 1($\tau = 17.23\%$), and Worst ($\tau = 40.21\%$) etc. And CIFAR-100N provides noisy labels, denoted as Fine ($\tau = 40.20\%$). To further discover the performance under different size of training set, we randomly select training samples from CIFAR-10N dataset, where the noise rate barely changes (variation range $< 1\%$), as shown in Table 2. Clothing1M is a large-scale clothing dataset with real-world noisy labels, whose images are crawled from several online shopping websites and labels are generated from the surrounding texts. In our setting, we use 1M training images with noisy label for training and 10K test images with clean label for evaluation. WebVision contains 2.4 million images crawled from the web whose training set contains many real-world noisy labels. Then following previous work (Li et al., 2020; Chen et al., 2019), baseline methods are compared on the first 50 classes of the Google image subset (approximately 66 thousand images). Models trained on mini WebVision are evaluated on both WebVision and ImageNet ILSVRC12 validation set.

Baselines: As introduced before, existing methods consist of loss correction and sample selection methods. To make comprehensive comparison, We select state-of-the-art methods from different categories. Specifically, (1) CE (Standard): standard training with cross-entropy loss. (2) T-Revision (Xia et al., 2019): loss correction with label transition matrix estimated without anchor points. (3) PTD (Xia et al., 2020b): estimate part-dependent label noise based on non-negative matrix factorization (NMF). (4) ELR+ (Liu et al., 2020): prevent memorization of the false labels by early-learning regularization. (5) DivideMix (Li et al., 2020): uses data division and MixMatch (Berthelot et al., 2019) mechanism to leverage corrupt data. (6) SOP (Liu et al., 2022): model label noise via another sparse over-parameterization and separate the underlying corruptions by exploiting implicit algorithmic regularizations .

Network structure and parameters: Our method is implemented by PyTorch v1.12. Baseline methods are implemented by the public codes with the same hyper-parameter settings in the original papers. For the methods with multiple networks, final test accuracy is evaluated with the average among the networks. To guarantee the comparability, we use the same backbone networks among the baseline methods for the same task.

For the CIFAR-10/100N datasets, we use 18/34-layer PreAct ResNet as backbone (He et al., 2016), following Li et al. (2020). The networks are trained for 300 epochs and SGD is used with a mo-

Table 2: Test accuracy (%) with different training samples on CIFAR-10N. The training data are randomly sampled from CIFAR-10N Worst set, with balanced categories and noise rate $\tau = 40\% \pm 1\%$. N is the number of training data. **Bold** means the highest reported accuracy and underline is the second highest accuracy.

Methods	CIFAR-10N ($\tau \approx 40\%$)						
	$N = 500$	2000	5000	10000	20000	40000	50000
CE (Standard)	32.54	41.05	49.58	58.61	63.84	74.33	76.86
T-Revision	28.54	29.64	32.69	63.47	77.37	80.66	82.10
PTD	18.99	26.59	39.01	65.85	66.69	70.97	80.16
ELR+	38.39	56.29	67.24	75.26	84.30	89.77	90.89
DivideMix	36.52	58.43	70.03	77.77	87.38	91.83	92.71
SOP	37.21	54.68	67.43	75.15	85.52	91.88	93.24
ClusterMix(ours)	**41.51**	**62.12**	**74.27**	**83.59**	**89.44**	**92.63**	**93.47**

mentum of 0.9 and a weight decay of 0.0005. The initial learning rate is set as 0.02 and reduced by a factor of 10 after the 150-th epoch. The batch size is 128, clustering interval is 10 epochs, and K-Means is used as cluster method. 10 warm-up epochs are used for CIFAR-10 and 30 warm-up epochs for CIFAR-100. For Clothing1M datasets, we use a 50-layer ResNet with ImageNet pre-trained weights. The networks are trained for 100 epochs and SGD with a momentum of 0.9, a weight decay of 0.001 is used. The initial learning rate is set as 0.002 and reduced by a factor of 10 after the 50-th epoch. The batch size is 32, clustering interval is 10 epochs, K-Means is used as cluster method, and 5 warm-up epochs are used. As Clothing1M is a large dataset with 1 million images, we randomly select 1000 batches with in each epoch. For (mini) WebVision dataset, we use the inception-resnet v2 (Szegedy et al., 2017), following the previous work. The networks are trained for 100 epochs and SGD with a momentum of 0.9, a weight decay of 0.001 is used. The initial learning rate is set as 0.01 and reduced by a factor of 10 after the 50-th epoch. The batch size is 32, clustering interval is 10 epochs, K-Means is used as cluster method, and 5 warm-up epochs are used. More parameters keep same with Li et al. (2020).

4.2 Classification Accuracy Evaluation

In this section, experiment results are shown on original CIFAR-N datasets, CIFAR-N datasets with different training samples, CIFAR-10 dataset with synthetic noise, clothing1M dataset, and WebVision dataset etc. More experiments are shown in the Appendix. Specifically, accuracy curve in Appendix D, experiments with synthetic label noise in Appendix F, sensitivity analysis for the number of clusters K and cluster methods in Appendix G, and contrast experiments for the sample selection module and the training module in Appendix H.

Experiment results on original CIFAR-N datasets. We first evaluate the test accuracy on the original CIFAR-N datasets, as shown in Table 1. Specifically, we select aggregate ($\tau = 9.03\%$), random 1 ($\tau = 17.23\%$), and Worst ($\tau = 40.21\%$) noisy label for CIFAR-10N and fine ($\tau = 40.20\%$) noisy label for CIFAR-100N. We report the averaged test accuracy over the last 10 epochs. In the original large dataset, our method reach state-of-the-art accuracy and outperforms slightly.

Experiment results on CIFAR-10 with different number of training samples. As the non-uniform noise would have greater impact with less training samples, we conduct experiments on CIFAR-10N Worst dataset with different number of training samples (500 –50000), as shown in Table 2. The samples are randomly sampled from the CIFAR-10N Worst dataset. The categories keep balance and the noise rate holds in $40\% \pm 1\%$. We report the averaged test accuracy over the last 10 epochs. ClusterMix outperforms state-of-the-art methods with various number of training samples. Notably, our method make greater improvement when training set goes smaller. The results demonstrate the effectiveness of our approach to suppress the inconsistent data space problem after sample selection, as the problem has greater impact when the training set is smaller.

Experiment results on Clothing1M and WebVision datasets. To validity the effectiveness of the proposed method on more general noisy datasets, experimental results on Clothing1M and (mini) WebVision datasets are shwon in Table 3. For Clothing1M, we design two experiments with different training samples: **Clothing1M-I**: Training with all 1 million samples available. **Clothing1M-II**: Training with randomly selected 5000 samples. The two tasks evaluate the performance of our

Table 3: Experimental results for Clothing1M and (mini) WebVision datasets. Clothing1M-I: Training with all 1 million samples of Clothing1M. Clothing1M-II: Training with randomly selected 5000 samples. The test accuracy (%) is evaluated on Clothing1M validation set, WebVision validation set, and ILSVRC12 validation set respectively. * means the result is copied from the original paper. **Bold** means the highest reported accuracy and underline is the second highest accuracy.

Methods	Clothing1M-I	Clothing1M-II	WebVision	ILSVRC12
CE (Standard)	69.55	45.11	-	-
T-Revision	74.18*	40.32	-	-
PTD	71.67*	25.33	-	-
ELR+	**74.81***	60.67	77.78*	70.29*
DivideMix	74.76*	56.57	77.32*	75.20*
SOP	73.50*	48.78	76.60*	69.10*
ClusterMix(ours)	74.34	**61.36**	**78.19**	**75.54**

Table 4: Test accuracy (%) on CIFAR-10 dataset with synthetic class-dependent non-uniform label noise. The class-dependent noise randomly selects five classes as low-noise classes (with noise rate τ_{low}) and the others as high-noise ones (τ_{high}). The second row of the table means the skewness (τ_{low}, τ_{high}) of noise rates. The generation details of the noisy labels are discussed in Appendix F. **Bold** means the highest reported accuracy and underline is the second highest accuracy.

Methods	CIFAR-10 (Average noise rate: 40%)				
	(0%, 80%)	(10%, 70%)	(20%, 60%)	(30%, 50%)	(40%, 40%)
CE (Standard)	63.93	66.83	70.98	73.95	78.99
T-Revision	69.57	73.01	76.72	81.03	84.11
PTD	69.73	70.93	75.86	78.98	82.15
ELR+	81.13	85.32	89.98	91.32	92.97
DivideMix	84.05	87.29	92.14	93.63	94.41
SOP	80.95	85.94	91.32	93.75	94.52
ClusterMix(ours)	**87.04**	**91.88**	**93.47**	**94.82**	**95.47**

method with different training data size. Although our method does not achieve the state-of-the-art accuracy on the very large dataset (in Task I), it outperforms all baseline methods when the number of samples reduces to a smaller quantity (in Task II). For (mini) WebVision dataset, we evaluate the models, which are trained on the mini WebVision training set, on WebVision validation set and ILSVRC12 validation set. Our method performs well on the both validation sets.

Experiments on CIFAR-10 dataset with synthetic label noises. To evaluate the performance under different skewness of the noise distribution, we conduct experiments on CIFAR-10 dataset with synthetic label noises. Table 4 shows the test accuracy for class-dependent label noise with varying skewness. Empirical results verify the effectiveness of the proposed method on the non-uniform noise. Besides, The results of symmetric and asymmetric noise are also available in Table 5, Appendix F.

5 Conclusion

Due to the memorization effect of DNNs (Arpit et al., 2017), the small-loss trick would select simple patterns (which usually have lower noise rate) first but regard all samples in the hard regions as corrupt data, resulting in inconsistent data space. Furthermore, the inconsistency problem would be harder when the training set is small and the label noise is heavy.

Therefore in this paper, we propose a novel ClusterMix algorithm to solve the inconsistent sample selection problem. ClusterMix combines the cluster-dependent sample selection method with a semi-supervised learning mechanism to distinguish and leverage the corrupt labels. Based on the idea of stratified sampling, our proposed cluster-dependent sample selection method would divide the dataset with variable criteria for samples in different clusters, which guarantees the sample selection from the entire sample space. Experiment results demonstrate that our method can suppress the inconsistent sample selection problem effectively and our ClusterMix outperforms all baseline methods on the datasets with a small number of training samples.

Ethics Statement. In this paper, our studies are not related to human subjects, practices to data set releases, potentially harmful insights, potential conflicts of interest and sponsorship, privacy and security issues, legal compliance, and research integrity issues. In real Scenarios, discrimination/bias/fairness may also result in the non-uniform annotated noisy labels. Due to the memorization effect of DNNs (Zhang et al., 2017), general training paradigms would also memorize the group bias of the annotators. Our proposed method makes a step for the robust and unbiased training with the non-uniform label noise.

Reproducibility. Experimental details are discussed in Section 4.1 and we will release the code upon acceptance.

REFERENCES

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 2017. URL http://proceedings.mlr.press/v70/arpit17a.html.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24392–24403, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html.

David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5050–5060, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1cd138d0499a68f4bb72bee04bbec2d7-Abstract.html.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 825–836. PMLR, 2021. URL http://proceedings.mlr.press/v139/berthon21a.html.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1062–1070. PMLR, 2019. URL http://proceedings.mlr.press/v97/chen19g.html.

Wenkai Chen, Chuang Zhu, and Yi Chen. Sample prior guided robust model learning to suppress noisy labels. *CoRR*, abs/2112.01197, 2021. URL https://arxiv.org/abs/2112.01197.

Jiacheng Cheng, Tongliang Liu, Kotagiri Ramamohanarao, and Dacheng Tao. Learning with bounded instance and label-dependent label noise. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1789–1799. PMLR, 2020. URL http://proceedings.mlr.press/v119/cheng20c.html.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks Learn. Syst.*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894. URL https://doi.org/10.1109/TNNLS.2013.2292894.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=H12GRgcxg.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi,

and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8536–8546, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a19744e268754fb0148b017647355b7b-Abstract.html.

Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W. Tsang, James T. Kwok, and Masashi Sugiyama. A survey of label-noise representation learning: Past, present and future. *CoRR*, abs/2011.04406, 2020. URL https://arxiv.org/abs/2011.04406.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, volume 9908 of *Lecture Notes in Computer Science*, pp. 630–645. Springer, 2016. doi: 10.1007/978-3-319-46493-0_38. URL https://doi.org/10.1007/978-3-319-46493-0_38.

Kurt Hornik, Maxwell B. Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8. URL https://doi.org/10.1016/0893-6080(89)90020-8.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2309–2318. PMLR, 2018. URL http://proceedings.mlr.press/v80/jiang18c.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017. URL http://arxiv.org/abs/1708.02862.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ea89621bee7c88b2c5be6681c8ef4906-Abstract.html.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14153–14172. PMLR, 2022. URL https://proceedings.mlr.press/v162/liu22w.html.

Yang Liu. Identifiability of label noise transition matrix. *CoRR*, abs/2202.02016, 2022. URL https://arxiv.org/abs/2202.02016.

Yang Liu and Jialu Wang. Can less be more? when increasing-to-balancing label noise rates considered beneficial. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 17467–17479, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/91e50fe1e39af2869d3336eaaeebdb43-Abstract.html.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL https://doi.org/10.1109/CVPR.2017.240.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5739–5748. PMLR, 2019. URL http://proceedings.mlr.press/v97/shen19e.html.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–19, 2022. doi: 10.1109/TNNLS.2022.3152527.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Satinder Singh and Shaul Markovitch (eds.), *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 4278–4284. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14806.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6835–6846, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/9308b0d6e5898366a4a986bc33f3d3e7-Abstract.html.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Jiankang Deng, Jiatong Li, and Yinian Mao. Extended T: learning with mixed closed-set and open-set noisy labels. *CoRR*, abs/2012.00932, 2020a. URL https://arxiv.org/abs/2012.00932.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/5607fe8879e4fd269e88387e8cb30b7e-Abstract.html.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2691–2699. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298885. URL https://doi.org/10.1109/CVPR.2015.7298885.

Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Maarten de Rijke, Qingyao Ai, Yufei Huang, Min Zhang, and Shaoping Ma. Improving web image search with contextual information. In Wenwu Zhu, Dacheng Tao, Xueqi Cheng, Peng Cui, Elke A. Rundensteiner, David Carmel, Qi He, and Jeffrey Xu Yu (eds.), *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pp. 1683–1692. ACM, 2019. doi: 10.1145/3357384.3358011. URL https://doi.org/10.1145/3357384.3358011.

Yilun Xu, Peng Cao, Yuqing Kong, and Yizhou Wang. L_dmi: A novel information-theoretic loss function for training deep nets robust to label noise. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 6222–6233, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8a1ee9f2b7abe6e88d1a479ab6a42c5e-Abstract.html.

Cheng Xue, Lequan Yu, Pengfei Chen, Qi Dou, and Pheng-Ann Heng. Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Trans. Medical Imaging*, 41(6):1371–1382, 2022. doi: 10.1109/TMI.2021.3140140. URL https://doi.org/10.1109/TMI.2021.3140140.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent label-noise transition matrix using dnns. *CoRR*, abs/2105.13001, 2021. URL https://arxiv.org/abs/2105.13001.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. Dual T: reducing estimation error for transition matrix in label-noise learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/512c5cad6c37edb98ae91c8a76c3a291-Abstract.html.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7164–7173. PMLR, 2019. URL http://proceedings.mlr.press/v97/yu19b.html.

Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao. Learning with biased complementary labels. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*, pp. 69–85. Springer, 2018. doi: 10.1007/978-3-030-01246-5_5. URL https://doi.org/10.1007/978-3-030-01246-5_5.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

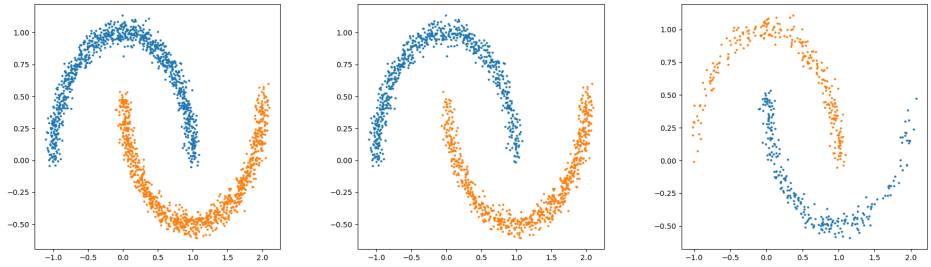
Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ZPa2SyGcbwh.

Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12912–12923. PMLR, 2021. URL http://proceedings.mlr.press/v139/zhu21e.html.

A EXPERIMENT RESULTS ON MOON DATA OF OUR CLUSTER-DEPENDENT SAMPLE SELECTION

As shown in Figure 1, existing methods cannot solve the classification problem of moon data on non-uniform label noise. Figure 5 shows the classification result and sample selection result of our cluster-dependent sample selection. Based on the idea of stratified sampling, we select clean samples from each cluster separately, using the small-loss trick and clean-label-dominate-the-cluster trick. The clean-label-dominate-the-cluster trick means that there are more cleanly labeled data in each cluster.

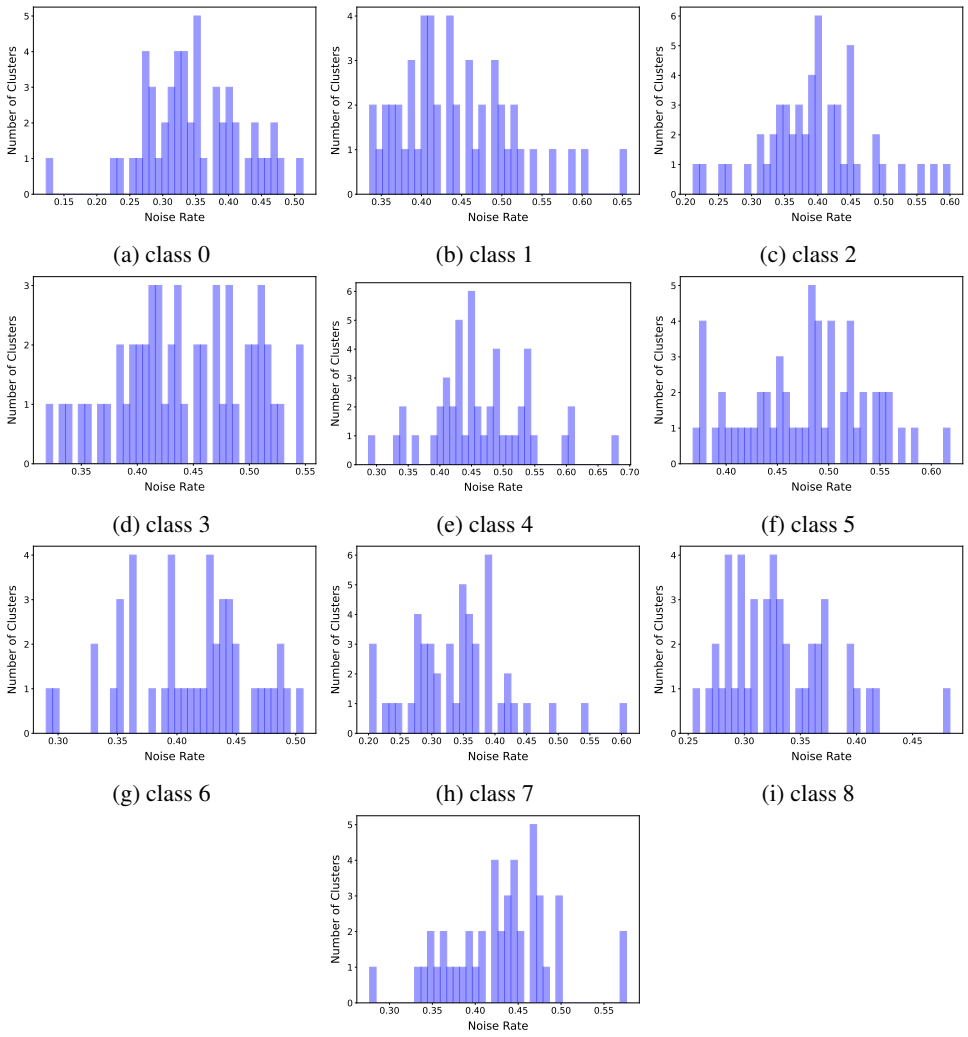


(a) Classification result (b) Clean data (c) Corrupt data

Figure 5: Experimental results on moon data of our cluster-dependent sample selection mechanism.

B NOISE RATE DISTRIBUTION ON EACH CLASS

As a supplementary of Figure 2, Figure 6 shows the noise rate varies not only on the entire dataset but also on each separate class. The results in Figure 6 illustrates the reality of our cluster-dependent sample selection rather than class -dependent sample selection.



(a) class 0 (b) class 1 (c) class 2

(d) class 3 (e) class 4 (f) class 5

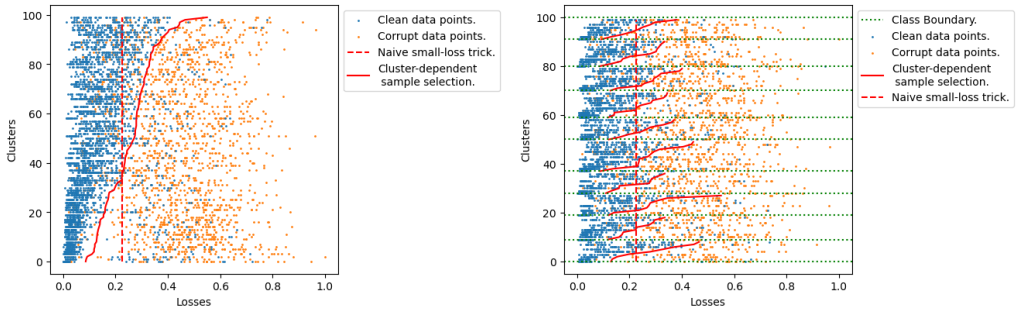
(g) class 6 (h) class 7 (i) class 8

(j) class 9

Figure 6: CIFAR-10 N Worst noise rate distribution on each class.

C ILLUSTRATION OF THE PROPOSED SAMPLE SELECTION STRATEGY

Figure 7 illustrates the effectiveness of the proposed cluster-dependent sample selection strategy. In the process of the small-loss samples selection, the proposed method adds another dimension and thus have variable criteria for different clusters. The proposed method can select more precise and adequate clean samples and thus achieve better evaluation accuracy. Besides, Figure 7b also shows the variability of the noise distribution inside each class.



(a) Clusters are sorted by the selection criteria.

(b) Clusters are first grouped by the main classes and then sorted by the selection criteria.

Figure 7: Comparison of the proposed cluster-dependent sample selection strategy and the naive sample selection. The red lines are the decision lines for the clean sample selection, i.e. The samples on the left of the lines will be selected as clean samples in the algorithms. In Figure 7a, the clusters are sorted by the selection criteria of the proposed cluster-dependent strategy. And in Figure 7b, the clusters are first sorted by main class index (The clusters between the adjacent green lines) and then the selection criteria of the proposed method.

D ACCURACY CURVE

Figure 8 shows the accuracy curve and standard deviation of the methods with different number of training samples. Our method performs slightly high than the baseline methods with large number of training samples. And the gap goes larger when the training samples are inadequate.

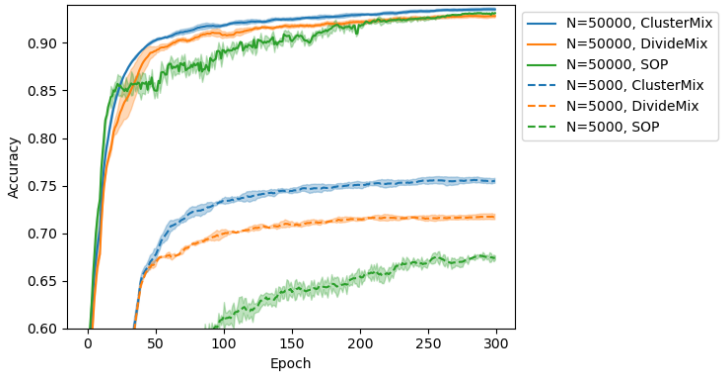


Figure 8: Accuracy curve on CIFAR-10N Worst dataset. ClusterMix (ours), DivideMix (baseline), and SOP (baseline) are shown in the figure. N denotes the number of training samples. The lines shows the mean values and shadows represent the std values with 5 repeats for each setting.

E EXAMPLE IMAGES OF CLUSTERS ON CIFAR-10 DATASET

Figure 9 shows several bird images of two clusters. The images in the same cluster share similar patterns and thus have similar corrupt probabilities. Conversely, the images in different clusters have far different patterns and have very different noise rate. The visualization of the real clusters further verify the validity of the proposed non-uniform noise model.

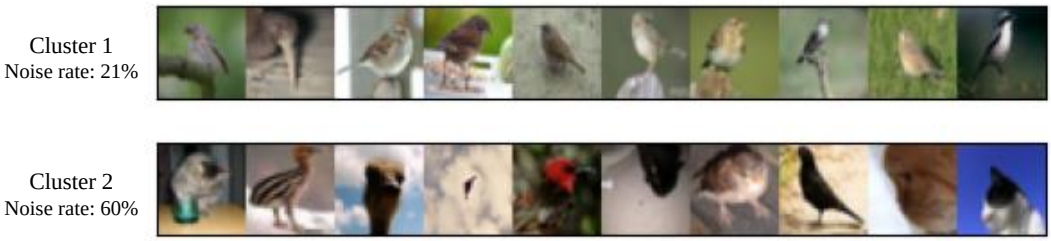


Figure 9: Example images of two bird clusters from CIFAR-10 dataset. The images in cluster 1 have similar simple patterns and would share similar lower corrupt probabilities. The images in cluster 2 have hard patterns and would have higher noise rate.

F EXPERIMENTS WITH SYNTHETIC LABEL NOISE

F.1 GENERATION OF THE SYNTHETIC NOISE

The generation of symmetric and asymmetric label noise is strictly following Li et al. (2020). Specifically, the symmetric label noise flips the labels in the uniform way, i.e. all the label share the same noise rate and flip to the other classes in a symmetric manner. The asymmetric label noise also flips labels with the same probabilities but just flip to the specific class for each class. The experiment results are shown in Table 5.

To further explore the methods with non-uniform noise, we synthesise a new class-dependent non-uniform noise on CIFAR-10 dataset. Specifically, we randomly select five classes as the low-noise-rate classes (noise rate is denoted as τ_{low}) and the other as the high-noise-rate ones (noise rate is denoted as τ_{high}). Then the noisy labels are generated with corresponding classes noise rates. The experiment results are shown in Table 4. In this experiments, we vary the skewness of the classes noise rate while keep the average noise rate the same. The experiment results are shown in Table 4.

F.2 RESULTS FOR THE SYMMETRIC AND ASYMMETRIC LABEL NOISE

Table 5 shows the experimental results on synthetic symmetric and asymmetric noise. From the experimental results on various synthetic noises (Tables 4 and 5), our proposed method can perform well under various distributional scenarios, especially when the noise rate is non-uniform and the noise is heavy. The empirical results verify the effectiveness of our method.

Table 5: Test accuracy (%) on CIFAR-10 dataset with synthetic symmetric and asymmetric label noise. **Bold** means the highest reported accuracy and underline is the second highest accuracy. * means the result is copied from the related papers.

Methods	CIFAR-10					
	Sym. 20%	Sym. 40%	Sym. 60%	Sym. 80%	Sym. 90%	Asy. (40%)
CE (Standard)	86.81	81.42	76.32	62.90	42.71	74.32
T-Revision	88.10	84.11	79.12	60.32	20.64	80.03
PTD	88.76	82.15	75.77	39.32	16.63	84.12
ELR+	94.60	93.58	92.82	91.10	75.20	92.07
DivideMix	95.97	94.82	93.40	**92.76**	75.40	92.05
SOP	93.18*	90.09*	86.76*	68.32*	67.78	91.43
ClusterMix(ours)	**96.35**	**95.50**	**95.36**	91.54	**77.94**	**92.63**

G SENSITIVITY ANALYSIS

In this section, we investigate the hyper-parameter sensitivity for the number of clusters K and cluster methods, respectively. For convenience, all the experiments in this section are conducted on the CIFAR-10N Worst dataset with 5000 randomly selected training samples. The 5000 selected samples are same with experiments in Table 2. The test accuracy and training time are reported from the same NVIDIA GeForce RTX 3090 GPU.

Different number of clusters. We first evaluate the test accuracy and training time for different number of clusters in our ClusterMix. The results are shown in Table 6. Both the test accuracy and training time increase with the increase of the number of clusters K, except $K > 800$. The experiment illustrate the effectiveness of stratified sampling. However, the training time and number of samples limit the number of clusters not too much. To balance the training effect and time, we prefer to select a intermediate value, which is 200 clusters (for 5000 samples) in the previous experiments.

Table 6: Test accuracy (%) and training time (h) with different number of clusters. The experiments are conducted on CIFAR-10N Worst dataset with 5000 training samples.

#Clusters K	1	5	10	50	100	200	500	800	1000
Accuracy	69.78	70.35	71.02	72.85	73.30	74.27	75.26	76.03	74.37
Training time	2.2h	2.3h	2.4h	2.6h	2.7h	3.0h	4.5h	5.0h	5.5h

Different clustering methods. Second, we analyse the sensitivity for the choices of different clustering methods. The results are shown in Table 7. Specifically, we compare the test accuracy and training time of four different clustering methods in our ClusterMix. K-Means, K-Means++, and Ward Hierarchical clustering require a specified number of cluster centers, which is set as 100 in the experiment. In contrast, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) can determine the number of clusters by itself. The experiment results demonstrate that our method achieves similar test accuracy and training time among different clustering algorithms, except DBSCAN. Therefore, we adopt K-Means as our clustering algorithm in all other experiments for convenience.

Table 7: Test accuracy (%) and training time (h) with different clustering methods. The experiments are conducted on CIFAR-10N Worst dataset with 5000 training samples.

Methods	K-Means	K-Means++	Hierarchical clustering	DBSCAN
Accuracy	74.27	74.10	74.25	70.84
Training time	3h	3h	3h	2.5h

H CONTRAST EXPERIMENTS FOR THE SAMPLE SELECTION MODULE AND THE TRAINING MODULE

To evaluate the effectiveness and applicability of the proposed cluster-dependent sample selection method, we combine the basic small-loss trick/cluster-dependent sample selection with two different training mechanisms (i.e. supervised learning and semi-supervised learning), as shown in Table 8. The experimental results validate the effectiveness and applicability of the proposed method.

- Effectiveness. Compared with basic small-loss trick, the proposed cluster-based sample selection strategy can improve the performance both in the supervised and semi-supervised mechanisms. Besides, the proposed method plays a greater role when the training samples are inadequate.
- Applicability. The proposed cluster-based sample selection strategy can be used to two label noise learning methods (one is supervised and the other is semi-supervised). Both of methods can improve the performance over the basic small-loss trick.

Table 8: Contrast experiments for the sample selection module and the training module. SL: supervised learning, SSL: semi-supervised learning. All the experiments are conducted on CIFAR-10N Worst dataset, as discussed in Table 8. N means the number of training samples. The first column of the table shows the sample selection methods and the second row shows the training mechanisms. And the number values are the experiment results of the combination of corresponding sample selection methods and training mechanisms.

	$N = 5000$		$N = 50000$	
	SL	SSL	SL	SSL
no selection	49.58	-	76.86	-
small-loss trick	66.41	70.03	84.47	92.71
cluster-dependent (ours)	71.96	74.27	88.53	93.47