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

Mengtian Zhang, Bo Jiang , Yuye Ling*, Xinbing Wang*

Shanghai Jiao Tong University
{zhangmengtian, bjiang, yuye.ling, xwang8}@sjtu.edu.cn

ABSTRACT

Learning with noisy labels is a challenging task in machine learning. Furthermore in reality, label noise can be highly non-uniform in 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. Inspired by stratified sampling, we propose a cluster-dependent sample selection algorithm followed by a contrastive training mechanism based on the cluster-dependent label noise. 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. The code is released at *https://github.com/MattZ-99/ClusterCL*.

***Index Terms*—** Non-uniform label noise, cluster dependent sample selection, contrastive training mechanism.

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 [1] and crowdsourcing [2] are cheap alternatives, which would produce datasets with noisy labels. Song et al. [3] reports that the overall ratio of corrupted labels in real-world datasets range from 8.0% to 38.5%.

Due to the universal approximation ability of DNNs, they can easily memorize and eventually overfit to the corrupted labels, leading to poor generalization [4]. Such overfitting would be aggravated by inadequate training samples, which often occur in real-world scenarios such as medical image processing. Efforts have been taken to robust learning paradigms under noisy labels [3]. Generally, existing methods on learning with noisy labels can be categorized into two groups: loss correction methods [5, 6, 7, 8, 9, 10] and sample selection methods [11, 12, 13, 14].

Methods in the first group mainly model label noise with label transition matrix. Several works [6, 7] assumed that label noises were class-dependent (called class-dependent noise, CDN). Other works [9, 10] proposed to model instance-dependent noise (IDN), which is more fine-grained but requires a large number of parameters to be estimated [16].

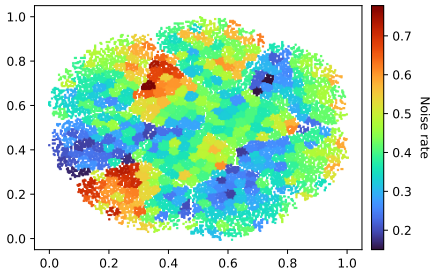


Fig. 1: Noise rate distribution of CIFAR-10N Worst dataset [15]. Each point represents a data sample in 2D t-SNE feature space. The color indicates the neighborhood noise rate around each data point.

Methods in the second group are designed to select confident clean samples from noisy datasets based on the memorization effect of DNNs [17], which tend to learn simple patterns first before fitting the corrupt samples. Han et al. [11] and Yu et al. [12] train two networks simultaneously where each network selects small-loss samples to train the other one. Furthermore, semi-supervised learning was used to explore both confident clean samples (as labeled data) and corrupt samples (as unlabeled data) [13, 14].

However, existing works explicitly or implicitly rely on the assumption of uniform noise rate over the entire dataset, while the local noise rates inside the real dataset can vary greatly, e.g. from 25% to 70% in Figure 1. Besides, Figure 1 also shows the non-uniformity of intra-class noise rate in more details. The methods of CDN transition matrix assume that all samples in 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. With the implicit assumption of uniform noise rate, the small-loss trick [18] would select simple patterns first and regard most samples in the hard regions as corrupt data (though part of the samples have correct labels). As a result, the inconsistent sample distribution would make neural networks confused in hard regions and thus weaken the generalization performance. A visualization illustration is shown in Figure 3.

To address the above inconsistent sample selection problem, we propose ClusterCL, which selects good samples from each cluster separately, where the clusters are based on features trained by a comprehensive weakly supervised mechanism. To get better sample selection and more robust model training, feature extraction, clustering, sample selection, and weakly supervised training are repeated and refined iteratively. Inspired by the idea of stratified sampling, ClusterCL selects samples in both simple and hard regions, alleviating the inconsistency defect of existing sample selection methods. Experimental results on both synthetic and real-world datasets verify the effectiveness of ClusterCL, which outperforms all baseline methods, especially with small training set and high noise rate.

*Corresponding authors: Bo Jiang, Yuye Ling.

This work is supported in part by the National Natural Science Foundation of China under Grant No. 62072302, and the Fundamental Research Funds for the Central Universities (project number YG2022QN058).

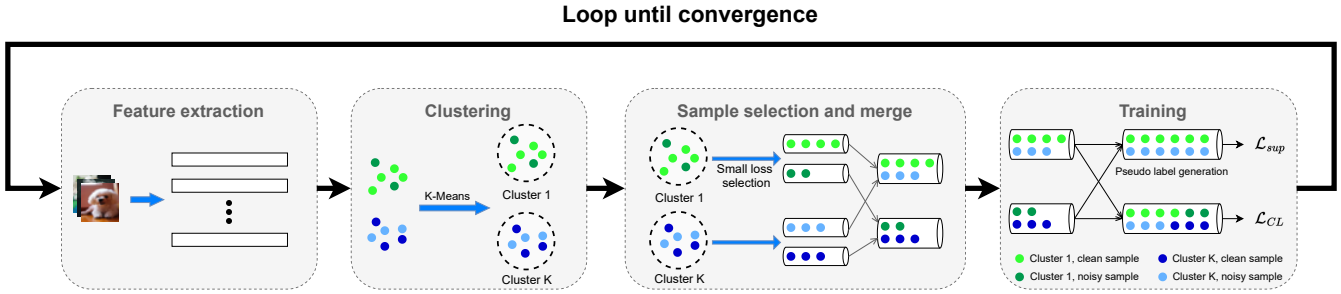


Fig. 2: An overview of proposed ClusterCL, which is robust to the non-uniform noise. Generally, the proposed method has four steps: 1) Feature extraction. Extract features using the current trained model. 2) Clustering. K-Means (or any other clustering method) is employed to group samples in the feature space. 3) Cluster-dependent sample selection. 2-dimension Gaussian Mixture Model (GMM) will be employed on the loss distribution and small-loss selection criteria are adopted to selection clean samples for each cluster separately. 4) Training for one epoch. Supervised cross-entropy loss $\mathcal{L}_{sup}$ and unsupervised contrastive loss $\mathcal{L}_{CL}$ will be used to update models' parameters. Note: the above four steps would be looped for epochs until convergence.

2. METHOD

An overview of framework is shown in Figure 2. Generally, our proposed method has four steps in each epoch, i.e. feature extraction, clustering, cluster-dependent sample selection and weakly supervised training. More details will be discussed in the following subsections.

2.1. Cluster-dependent Sample Selection (CDSS)

Clustering algorithm, cluster-dependent data division, and effectiveness analysis would be discussed in this section.

Clustering. To group the entire dataset by the similarity, we use the features extracted by backbone networks trained in previous epoch to divide the dataset into clusters. Experiments in Section 3.5 show the robustness of the proposed method to the clustering algorithm and clusters number. Therefore for convenience, K-means is selected as the clustering method unless otherwise specified.

Cluster-Dependent Data Division. As shown above, the samples in different clusters have different noise rate. To address the challenge of non-uniform label noise, we propose to select clean and corrupt samples within clusters.

Inspired by the idea of stratified sampling, we propose to first divide the entire dataset into several cluster-based subset. Then 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.

Effectiveness analysis. Two clusters with different noise rates are selected in Figure 3 to illustrate the effectiveness of proposed method in sample selection process. Specifically, different criteria would be adopted for different clusters, based on the fitting degree of neural networks. As shown in Figure 3, the baseline method (red line in Figure 3a) would omit most of samples in Cluster 1, which has higher noise rate and neural networks would fit it slowly. Therefore compared with baseline method, the proposed adaptive selection method sets variable criteria for different "sub-populations" (i.e. clusters with different fitting characteristics). The effectiveness of the proposed method would be summarized as:

1) For clusters with smaller losses (Cluster 0 in Figure 3), the proposed cluster-dependent sample selection method would achieve **higher precision** with **similar recall**.

2) For clusters with higher losses (Cluster 1 in Figure 3), the proposed method would achieve **much higher recall** with **similar precision**.

2.2. Contrastive Training Mechanism

The training process of our proposed method contains two parts, i.e. unsupervised training and supervised training. Unsupervised contrastive training is used to pull similar samples together which facilitates the cluster-dependent selection and supervised training is conducted using selected clean samples to push samples from different classes away.

Co-training mechanism. To avoid the accumulation of confirmation bias in self-training, we follow the common practice [11, 12] of co-training two networks simultaneously.

Noisy data usage. To fully utilize the entire dataset (both clean and corrupt data), the proposed method keeps the clean labels while replacing the corrupted ones by generated pseudo labels. Specifically, the pseudo labels are generated with aggregated output of two networks and multiple augmentation views [14].

Mixup. To regularize the neural network for more robust training, mixup [19] and FMix [20] are employed to construct more diverse sample views. Specifically, for a pair of samples $(x_1, \hat{y}_1)$ and $(x_2, \hat{y}_2)$, the mixed views are computed by

$$\begin{aligned}(x', y') &= \lambda\,(x_1, \hat{y}_1) + (1-\lambda)\,(x_2, \hat{y}_2)\\ (x'', y'') &= \text{MASK}\,(x_1, \hat{y}_1) + (1-\text{MASK})\,(x_2, \hat{y}_2)\end{aligned} \tag{1}$$

where $\lambda \sim \text{Beta}(\alpha = 4, \alpha = 4)$ and MASK means a sampled mask for the input data. Denoting by $\mathcal{L}_{CE}$ the cross-entropy loss, the supervised training loss $\mathcal{L}_{sup}$ can be written as

$$\begin{aligned}\mathcal{L}_{sup} &= \mathcal{L}_{mixup}(x', y') + \mathcal{L}_{fmix}(x'', y'')\\ &= \mathcal{L}_{CE}(x', y') + \mathcal{L}_{CE}(x'', y'')\end{aligned} \tag{2}$$

Contrastive loss. To pull similar samples together in the feature space, additional unsupervised contrastive loss is employed on the last-layer features (denoted as z_i) for the entire dataset. Specifically, with the similarity measured by dot product, the contrastive loss $\mathcal{L}_{CL}$ we use in this paper is given by

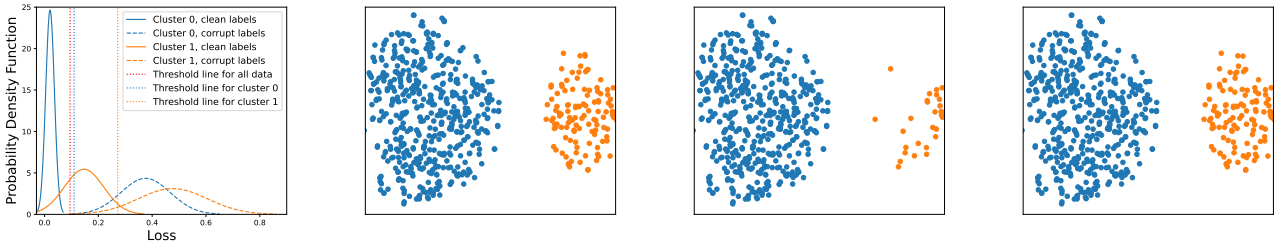
$$\mathcal{L}_{CL} = \sum_{i=1}^{N} l_{CL}(i) = -\sum_{i=1}^{N} \log \frac{\exp\left(z_i \cdot z_i^{+}/t\right)}{\sum_{j=1}^{K}(z_i \cdot z_j/t)}, \tag{3}$$

where N is the batch size, K is the feature dictionary size as in [21], and t is a temperature hyper-parameter.

Overall training loss. The overall loss in the training stage is given by

$$\mathcal{L}_{total} = \mathcal{L}_{sup} + \lambda \mathcal{L}_{CL}, \tag{4}$$

where λ is a weight factor, which is set to 0.1 in our experiments. Two components of the losses are employed to achieve inter-class separability [22] and intra-cluster invariance [23] respectively.



(a) Loss Distribution (b) Ground Truth (c) Baseline Method (d) CDSS (ours)

Fig. 3: Visualization comparison of clean sample selection results between proposed cluster-dependent method and previous baseline [13]. Color **blue** and **orange** are two selected clusters from CIFAR-10N Worst dataset [15], with noisy rate 21.1% and 60.8% respectively. The loss distribution of clean and corrupt samples in the two clusters are shown in Figure 3a and the vertical lines are the selection criteria. Figure 3c, Figure 3c, and Figure 3d respectively display the selected clean samples of ground truth, baseline method, and ours cluster-dependent sample selection (CDSS) with 2D T-SNE visualization.

3. EXPERIMENTS

3.1. Datasets and Setup

Datasets: We evaluate our method on three real noisy dataset CIFAR-10N, CIFAR-100N [15], and Clothing1M [24].

Baselines: To make comprehensive comparison, We select state-of-the-art methods from different categories: CE (standard training with cross-entropy loss), T-Revision [6], PTD [9], ELR+ [25], DivideMix [13], SOP [26], and ProMix [27].

Network structure and parameters: 18/34-layer PreAct ResNet and 50-layer ResNet [28] with ImageNet pretrained weights are used as backbone for CIFAR-10/100N and Clothing1M. As Clothing1M is a large dataset with 1 million images, we randomly select 1000 batches with in each epoch. To guarantee the comparability, we use the same backbone networks among the baseline methods for the same task.

3.2. Classification Accuracy Evaluation

Experiment results on CIFAR-10/100N. Experimental results on CIFAR-N datasets are shown in Table 1. The reported accuracy is averaged over the last 10 epochs. In the original dataset with 50,000 training samples, our method can outperforms all the baseline methods. Besides as discussed in Table 4, our method can achieve better performance with inadequate samples.

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 $\tau = 40.20\%$.

Methods	CIFAR-10N			CIFAR-100N
	Aggr	Rand1	Worst	Fine
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
ProMix	96.83	96.17	94.05	70.54
SOP	95.61	95.28	93.24	67.81
ClusterCL(ours)	**96.86**	**96.29**	**94.13**	**71.87**

Experiment results on Clothing1M datasets. To validity the effectiveness of the proposed method on more general noisy datasets, experimental results on Clothing1M are shwon in Table 2. 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 assess the performance of our method using varying training set sizes. Our method achieves the state-of-the-art accuracy on the more realistic dataset, particularly when the sample size is limited.

Table 2: Experimental results for Clothing1M. * means the result is copied from the original paper.

Methods	Clothing1M-I	Clothing1M-II
CE (Standard)	69.55	45.11
T-Revision	74.18*	40.32
PTD	71.67*	25.33
ELR+	74.81*	60.67
DivideMix	74.76*	56.57
ProMix	72.85	55.39
SOP	73.50*	48.78
ClusterCL(ours)	**74.84**	**61.98**

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 varying training set sizes(500 –50000), as shown in Table 4. ClusterCL 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 distribution problem after sample selection, as the problem has greater impact when the training set is smaller.

Table 3: Precision (%), Recall (%), and F1-score (%) in the clean sample selection step on CIFAR-10 Worst dataset with 5000 training samples. For the standard Cross-Entropy method, all the samples are regraded as the clean ones.

Methods	Precision	Recall	F1-score
CE (Standard)	59.02	100.0	74.22
DivideMix	88.34	82.48	85.31
ProMix	**89.19**	82.03	85.46
ClusterCL(ours)	89.04	**88.45**	**88.74**

3.3. Sample Selection Evaluation

In this section, we would provide empirical evaluation of the clean sample selection process with several baseline methods. The experi-

Table 4: 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.

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
ProMix	34.96	58.15	69.75	77.95	88.06	92.71	94.05
SOP	37.21	54.68	67.43	75.15	85.52	91.88	93.24
ClusterCL(ours)	**44.26**	**64.31**	**76.19**	**84.67**	**89.85**	**93.08**	**94.13**

Table 5: Ablation study for the proposed method. The experiments are performed on CIFAR-10N Worst dataset [15] with 40% noise rate. CDSS represents proposed cluster-dependent sample selection and GMM represents Gaussian mixture module used in baseline method DivideMix [13]. CE means cross-entropy loss and CL means contrastive loss.

Sample selection	-	CDSS	CDSS	GMM	CDSS	CDSS	CDSS
Mixup	-	-	-	Mixup	Mixup	Mixup+FMix	Mixup+FMix
Number of nets	single	single	dual	dual	dual	dual	dual
Losses	CE	CE	CE	CE	CE	CE	CE+CL
Accuracy	76.86	90.48	92.15	92.71	93.47	93.60	94.13 (ours)

mental results are shown in Table 3. Our proposed cluster-dependent sample selection method would achieve slightly lower precision but much higher recall and thus higher F1-score. The experiment illustrates the effectiveness of the proposed method.

3.4. Ablation Study

Ablation study for the proposed method is shown in Table 5. The experiments are performed on CIFAR-10N Worst dataset [15] with 40% noise rate and 50,000 training samples. The experimental results prove the effectiveness of the proposed cluster-dependent sample selection mechanism, mixup regularization, co-training mechanism, CE and CL losses. The proposed cluster-dependent sample selection mechanism performs a foundational role. Besides, the weakly-supervised technologies, including mixup, co-training, and contrastive learning enhance the final performance.

3.5. Sensitivity Analysis

Experiments in Tables 6 and 7 shows the sensitivity of the proposed method to clustering algorithm and numbers. The experiments are conducted on CIFAR-10N Worst dataset with 5000 training samples.

Experiments in Table 6 shows the method with **different number of clusters.** 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: Experiments with different number of clusters.

#Clusters K	1	200	500	800	1000
Accuracy	69.78	74.27	75.26	76.03	74.37
Time	2.2h	3h	4.5h	5.0h	5.5h

Experiments in Table 7 shows the method with **different clustering methods.** The proposed method achieves similar test accuracy and training time among different clustering algorithms, except DBSCAN. Therefore, K-Means is adopted as our clustering algorithm in all other experiments for convenience.

Table 7: Experiments with different clustering methods.

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

4. CONCLUSION

Due to the memorization effect of DNNs [17], 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 distribution. Furthermore, the inconsistency problem would be harder when the training set is small and the label noise is heavy.

Inspired by the idea of stratified sampling, the paper proposed a novel ClusterCL mechanism to solve the robust learning challenge under non-uniform label noise. ClusterCL combines the cluster-dependent sample selection method with a weakly supervised learning mechanism to distinguish and leverage the noisy labels. At training stage, supervised cross-entropy loss and unsupervised contrastive loss are employed together to achieve inter-class separability and intra-cluster invariance in feature space. At sample selection stage, the dataset would be divided into clusters and small-loss selection would be performed on each cluster respectively. The proposed cluster-dependent sample selection method sets adaptive criteria for samples with different fitting degree and thus brings a significant improvement on selection recall.

Experiment results demonstrate that the proposed ClusterCL can suppress the inconsistent sample selection problem effectively and thus outperforms all baseline methods on the datasets, especially when the training set is small and noise rate is high.

5. REFERENCES

[1] 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 *CIKM*, 2019, pp. 1683–1692.

[2] Xiyu Yu, Tongliang Liu, Mingming Gong, and Dacheng Tao, "Learning with biased complementary labels," in *ECCV*. 2018, vol. 11205, pp. 69–85, Springer.

[3] 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.

[4] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization," in *ICLR*, 2017.

[5] 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 *CVPR*, 2017, pp. 2233–2241.

[6] 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 *NeurIPS*, 2019, pp. 6835–6846.

[7] 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 *NeurIPS*, 2020.

[8] 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," *NeurIPS*, vol. 32, 2019.

[9] 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 *NeurIPS*, 2020.

[10] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama, "Confidence scores make instance-dependent label-noise learning possible," in *ICML*. 2021, vol. 139, pp. 825–836, PMLR.

[11] 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 *NeurIPS*, 2018, pp. 8536–8546.

[12] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama, "How does disagreement help generalization against label corruption?," in *ICML*. PMLR, 2019, pp. 7164–7173.

[13] Junnan Li, Richard Socher, and Steven C. H. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR*, 2020.

[14] 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 *NeurIPS*, 2021, pp. 24392–24403.

[15] 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 *ICLR*, 2022.

[16] Yang Liu, "Identifiability of label noise transition matrix," *CoRR*, vol. abs/2202.02016, 2022.

[17] 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 *ICML*. PMLR, 2017, pp. 233–242.

[18] Wenkai Chen, Chuang Zhu, and Yi Chen, "Sample prior guided robust model learning to suppress noisy labels," *CoRR*, vol. abs/2112.01197, 2021.

[19] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[20] Ethan Harris, Antonia Marcu, Matthew Painter, Mahesan Niranjan, Adam Prügel-Bennett, and Jonathon S. Hare, "Understanding and enhancing mixed sample data augmentation," *CoRR*, vol. abs/2002.12047, 2020.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[22] Rudrajit Das and Subhasis Chaudhuri, "On the separability of classes with the cross-entropy loss function," *CoRR*, vol. abs/1909.06930, 2019.

[23] Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman, "Investigating why contrastive learning benefits robustness against label noise," in *ICML*. PMLR, 2022, pp. 24851–24871.

[24] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang, "Learning from massive noisy labeled data for image classification," in *CVPR*. 2015, pp. 2691–2699, IEEE Computer Society.

[25] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," in *NeurIPS*, 2020.

[26] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You, "Robust training under label noise by over-parameterization," in *ICML*. 2022, vol. 162, pp. 14153–14172, PMLR.

[27] Haobo Wang, Ruixuan Xiao, Yiwen Dong, Lei Feng, and Junbo Zhao, "Promix: Combating label noise via maximizing clean sample utility," *CoRR*, vol. abs/2207.10276, 2022.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *ECCV*. 2016, vol. 9908 of *Lecture Notes in Computer Science*, pp. 630–645, Springer.