Improving Training For Noisy Labels

Anuj Gupta
Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
21D070014@iitb.ac.in

***Abstract*—Machine Learning has achieved remarkable success, consuming vast amounts of data. This creates an alarming need for clean data for accurate training, but real-world data always contains some noise. Thus, robust training on noisy datasets is gaining momentum each day. This report proposes a novel pipeline that leverages self-supervised learning (SimCLR), iterative pseudo-label refinement and a newly proposed Softmax-weighted Cross-entropy loss function to enhance model robustness against instance-dependent noise (IDN). Experiments conducted on CIFAR10, CIFAR100 and other medical datasets like Chest X-ray and Chaoyang demonstrate that the proposed methodology significantly improves performance compared to several previous methods under varying noise levels.**

I. Introduction: Noisy Label Problem

The noisy label problem is when the labels assigned to data points in a dataset are incorrect or unreliable. In other words, some training data has been mislabeled due to human error during the labelling process or inherent ambiguity in the data itself. It addresses the challenge of training DNNs with noisy labels in real-world scenarios, emphasising the need to mitigate adverse effects caused by corrupted labels. The goal is to enable noise-tolerant training for deep learning to prevent degradation in generalisations on unseen data [1].

A. Causes of Noisy Labels

- **Incorrect Labels:** Some data points may be labelled with the wrong class label or annotation. For example, an image of a cat might be labelled as a dog, or a positive review might be labelled as negative.
- **Ambiguous Labels:** In this case, the true label of a data point may be subjective or ambiguous. For instance, determining whether a product review is positive or negative might depend on individual interpretation.

B. Impacts of Noisy Labels

The extent of label noise can vary from dataset to dataset. In some cases, only a small percentage of labels may be noisy, while in others, a significant portion may be incorrect [1].

- **Decreased Model Performance:** Noisy labels lead to decreased model performance as the model learns from incorrect or misleading examples. For instance, in a dataset, some images labelled as cats might contain dogs, and vice versa. If the model learns from these mislabeled examples, it may generalise poorly to unseen data.
- **Bias in Training:** Noisy labels can introduce bias into the trained models, especially if certain classes are more prone to mislabeling than others. This can lead to unfair predictions. We may be prone to classify a fox as a dog rather than a lion as a dog.
- **Overfitting:** Noisy labels can cause the model to overfit the training data. Instead of identifying the general underlying pattern, it may memorise the noisy labels, making it difficult for the model to predict unseen data.

C. Challenges in Noisy Labels

- **Insufficient Data:** If the noise rate is large, less clean data would remain for actual training, and identifying this clean data is challenging.
- **Increased Training Time and Resource Consumption:** Training with noisy labels may require more iterations and computational resources to perform satisfactorily, increasing resource consumption and training time.
- **Difficulty in Label Cleaning:** Identifying and correcting noisy labels is challenging and time-consuming. Manual inspection and correction of mislabeled examples can be impractical for large datasets, and automatic cleaning may induce errors.

II. Related Works

In machine learning, dealing with noisy labels has been a persistent challenge, as label noise often leads to overfitting and poor generalisation in deep neural networks (DNNs). Various approaches have been proposed to mitigate these issues, using traditional non-deep learning methods and advanced deep learning strategies.

A. Non-Deep Learning Approaches [1]

Traditional methods primarily focus on data cleaning, surrogate loss functions, and probabilistic models.

- **Data Cleaning:** Techniques such as bagging and boosting are employed to identify and exclude corrupted labels. However, these methods risk over-cleaning, potentially removing true-labelled examples.
- **Surrogate Loss Functions:** Researchers propose convex surrogate loss functions to handle computational challenges, especially in binary classification. Unfortunately, these are not generalisable to multiclass tasks.
- **Probabilistic Models:** These approaches estimate label confidence by clustering feature distributions, using this information for weighted training or converting hard

labels into soft labels. Bayesian methods enhance robustness but often increase overfitting due to their complexity.
- **Model-Based Methods:** Modifications to traditional models like Support Vector Machines (SVM) and decision trees improve robustness against label noise. However, these design principles are difficult to integrate with DNNs.

B. Deep Learning Approaches [1]

The advent of DNNs has led to more sophisticated methods for handling noisy labels:

- **Robust Architectures:** Adding noise adaptation layers to model label transition processes or creating architectures that better accommodate diverse noise types.
- **Robust Regularization:** Techniques that prevent overfitting to false-labeled examples.
- **Robust Loss Functions:** Modifications to standard loss functions to make them more resistant to noise.
- **Loss Adjustment:** Reweighting or correcting losses based on label confidence.
- **Sample Selection:** Identifying and focusing on true-labeled examples via multi-network or multi-round learning approaches.

The paper Normalized Loss Functions for Deep Learning with Noisy Labels [2] introduces the concept of normalising loss functions to improve robustness. However, normalisation can lead to underfitting. The Active-Passive loss framework is proposed to address this, combining active loss functions (e.g., Cross-Entropy) with passive loss functions (e.g., Mean Absolute Error) to prevent underfitting. This approach achieves state-of-the-art (SOTA) performance on datasets like CIFAR-10 and CIFAR-100 for symmetric and asymmetric noise. SELFIE: Refurbishing Unclean Samples for Robust Deep Learning [3] presents a hybrid method combining loss correction and sample selection. It classifies samples as clean or refurbishable based on predictive uncertainty and updates them iteratively to improve the training dataset quality. SELFIE outperforms other state-of-the-art techniques like Active Bias and Co-teaching for pair and symmetric noise on CIFAR-10, CIFAR-100 and Tiny-ImageNet when trained using DenseNet. Several other approaches have explored unique methods to address noisy labels. The reverse k-NN Algorithm used in [4] introduces structural labels based on the Reverse k-NN algorithm to better capture the feature distribution of data. It extracts distribution information using reverse k-NN and integrates it into training to reduce reliance on noisy predictions. Additionally, it applies strong augmentations and mixup strategies during training to avoid overfitting to noisy samples. The proposed method is tested on benchmarks like CIFAR-10, CIFAR-100, and WebVision, achieving state-of-the-art performance under noisy label conditions by improving the feature manifold and generalisation abilities.

Active Bias [5] focuses on high-variance samples to improve generalisation. Open-Set Label Noise [6] addresses learning with noisy labels in weakly supervised learning, proposing Open-set samples with Dynamic Noisy Labels (ODNL). This regularisation technique utilises open-set noises to enhance model robustness without disrupting learning from clean data. Experimental results validate ODNL's effectiveness in improving robust algorithms and enhancing Out-of-Distribution detection tasks, even in noisy label settings. Adaptive Robust Loss Functions [7] presents a generalised loss function, incorporating robustness as a continuous parameter, and it encompasses various loss functions that enable automatic adjustment of loss robustness during neural network training. DivideMix [8] introduces a dynamic approach for handling noisy labels by modelling the per-sample loss distribution with a mixture model. This allows the training data to be divided into a labelled set containing clean samples and an unlabeled set with noisy samples. The method then employs semi-supervised training on both sets, incorporating enhanced MixMatch strategies, including label co-refinement for labelled samples and label co-guessing for unlabeled samples. [16] addresses the problem of underfitting in deep learning with noisy labels by proposing a new framework that uses label confidence to prioritize clean samples during training. Based on robust loss methods, this approach selectively suppresses noisy data without harming clean data learning. The paper proved that its method can still reach the robust loss optimum under certain conditions. It showed through experiments on synthetic (CIFAR datasets) and real-world datasets (Clothing and WebVision) that its framework outperforms state-of-the-art techniques.

C. Unexplored Area: Self-Supervised Learning

Despite the extensive efforts in robust architectures, loss functions, and hybrid approaches, none of these methods has explored *self-supervised learning frameworks* for handling noisy labels and datasets with Instance-dependent Noise (IDN). Self-supervised methods like SimCLR and BYOL, which leverage unlabelled data for pretext tasks, have shown promise in improving model robustness and generalisation. Our approach uniquely integrates self-supervised learning on IDN datasets to address label noise along with consensus-based pseudo-labelling, representing a significant step forward in this research domain.

III. Proposed Method

Many proposed methods assume specific noise structures like symmetric or asymmetric noise, limiting their applicability in real-world datasets. Hence, we proposed a pipeline robust to Instance-dependent Noise (IDN), which closely resembles real-life noise. Label corruption in IDN depends on input features. Therefore, it represents a more realistic but challenging scenario.

Proposed Loss Functions

For the main training stage, we proposed a **Softmax-weighted Cross-Entropy Loss**. Specifically, given softmax probabilities $p_{i,c}$ for sample i over classes $c \in \mathcal{C}$, and the ground-truth label $y_i \in \{1, \ldots, C\}$, the per-sample cross-entropy loss is computed as

$$l_i = -\log(p_{i,y_i})$$

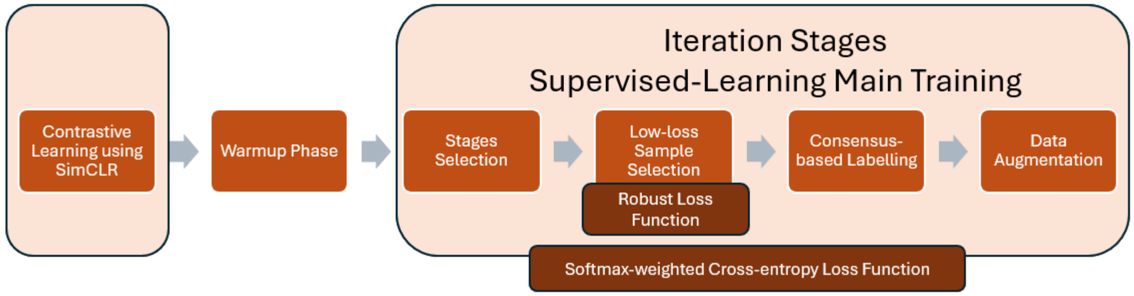

Fig. 1: Model Pipeline Flowchart

To emphasize samples where the model is more confident, we introduce a softmax-weighted formulation:

$$\begin{aligned} L &= \frac{1}{N}\sum_{i=1}^{N} p_{i,y_i} \times l_i \\ &= -\frac{1}{N}\sum_{i=1}^{N} p_{i,y_i} \times \log(p_{i,y_i}) \end{aligned}$$

where N denotes the batch size and L is our final loss function for main training.

Why it helps:

- **Using weighted loss function** prioritizes high-confidence samples while reducing noisy labels's influence.
- **Greater confidence for more certain samples increases loss**, ensuring the model learns more from reliable labels.
- **More reliable supervision** is achieved by emphasizing samples with correct labels, leading to stronger learning.
- **Robustness against noisy labels** is improved as uncertain samples contribute less, preventing error reinforcement.

For pseudo-label assignment and low-loss sample selection, we adopt a composite and robust active-passive loss combining **Normalized Cross-Entropy (NCE)** and **Mean Absolute Error (MAE)** [2]. For a sample with one-hot encoded label $q_{i,c}$, the selection loss is defined as:

$$\begin{aligned} L &= \text{NCE} + \text{MAE} \\ &= -\frac{\sum_{c=1}^{C} q_{i,c}\log(p_{i,c})}{\sum_{j=1}^{C}\sum_{c=1}^{C} q_{i,j}\log(p_{i,c})} + \sum_{c=1}^{C} |p_{i,c} - q_{i,c}| \end{aligned}$$

NCE balances the penalty across samples. It mitigates memorization of noise by normalizing loss magnitudes. Conversely, MAE is label-insensitive as it doesn't assume the label is always correct. It acts like distributional matching rather than forcing hard classification. Also it is known to be noise-tolerant and robust against outliers.

Why the Combination Works: Our pipeline has two different roles:

- Training: For this, we need stable gradients and a strong optimisation signal.
- Pseudo-label selection: We need to identify trustworthy labels, even when noisy.

Hence, the combination serves both purposes, where each part complements the other during the training.

A. Pipeline Developed

We propose a hybrid pipeline combining self-supervised learning (SSL) and iterative pseudo-label refinement to address the challenges of training on IDN datasets. Our methodology integrates SimCLR-based SSL pretraining, stage-wise filtering, consensus-based labelling, and dynamic data augmentation to robustly adapt to noisy labels.

SimCLR-based SSL Pretraining: To improve the feature representations of the model in noisy environments, we begin with contrastive learning using SimCLR. In this phase:

- The SimCLR framework uses contrastive learning, where augmented views of the same image are treated as positive pairs, and different images within the batch are treated as negative pairs.
- To create diverse positive pairs, we apply strong augmentations, including RandomResizedCrop, ColorJitter, GaussianBlur, and RandomGrayscale.
- This pretraining phase focuses solely on feature learning, independent of label noise, and equips the model with noise-resilient embeddings for downstream tasks.

Warmup Phase: After pretraining, the model undergoes a warmup phase (5 epochs), where it is trained on the noisy dataset using Softmax-weighted Cross-entropy Loss (SWCE). This initialises the model for supervised fine-tuning and ensures it captures basic patterns in the data.

Training with Stages: The training continues with four iterations of pseudo-label refinement, structured around specific stages:

- Stage Definition: Some epochs are marked in each iteration. We have called them stages. These stages are empirically selected (For example, epochs 2, 3 and 4 are marked as stages for iteration 1 with four epochs and epochs 2, 5 and 7 are marked for subsequent iterations of 7 epochs). At these stages, the model evaluates the training dataset.
- Loss Threshold Filtering: The model computes the loss for all data points during each stage. Samples with a loss below a predefined threshold are identified and stored. This ensures the selection of confidently predicted samples.

Pseudo-label Generation: At the end of each iteration, the samples identified across all stages are analysed:

- Consensus-based Labeling: Only samples consistently selected across all stages are pseudo-labeled using the model's predictions. These "common" samples are treated as clean data. This is by the assumption that before the network entirely fits the noisy labels, the label prediction of mislabeled samples either changes inconsistently or corresponds to accurate label [3].
- Preservation of Remaining Labels: The labels of other samples (those not consistently selected) are retained in their original state to prevent the amplification of label noise.

Dynamic Data Augmentation: For the pseudo-labeled samples, data augmentation is applied to enhance diversity. Augmented images of these samples, assigned the same pseudo-labels, are added to the training data. This step improves the model's generalisation and mitigates overfitting to noisy labels.

Training Continuation: The training resumes with updated labels and augmented data, repeating for subsequent iterations.

B. Key Innovation

- **SimCLR-based Pretraining for Noise Robustness**: Leveraging contrastive learning during pretraining helps the model learn noise-agnostic features, reducing the impact of label noise in downstream tasks.
- **Stage-Wise Iterative Refinement**: Dividing training into stages helps identify consistent samples over time, leading to more reliable pseudo-labeling and reducing the propagation of erroneous labels.
- **Handling Instance-Dependent Noise**: Unlike approaches that assume independent label noise, our method adapts to IDN through iterative refinement, which dynamically distinguishes between clean and noisy labels.
- **Augmented Pseudo-Labeling**: Combining pseudo-labeled samples with augmentations strengthens the model's ability to generalize clean data patterns, reducing the risk of overfitting to noisy information.
- **Confidence-weighted loss**: Introducing a confidence-weighted loss function that dynamically prioritizes reliable samples, enhancing pseudo-label refinement and robustness against noisy labels.

C. Empirical Configuration

In our experiments, the stages were 2, 3 and 4 in the first iteration and 2, 5 and 7 in the subsequent iterations. The loss threshold for filtering was set to 1 for CIFAR datasets and 0.25 for medical datasets, determined through empirical evaluations. SimCLR pretraining was conducted for 30 epochs with augmentations like RandomResizedCrop, ColorJitter, RandomGrayscale, and GaussianBlur.

IV. Experimental Setup

A. Dataset and Noise Injection

- We evaluated the performance on CIFAR and medical datasets. The dataset details are given in Table I.
- Instance-dependent noise was introduced at three levels: 20%, 30% and 50% for CIFAR datasets and 20% and 30% for medical datasets. The noise was generated such that the incorrect labels depended on the input features, simulating real-world scenarios.

Dataset	# of Classes	# of Training Samples	# of Validation + Test Samples
CIFAR10	10	50000	10000
CIFAR100	100	50000	10000
Chest X-ray	2	5216	640
Chaoyang	4	4021	2139

TABLE I: Summary of datasets used.

B. Model Architecture

The ResNet-18 architecture was employed for all experiments. This model balances computational efficiency and performance for different datasets. However, experiments were also performed on Resnet-34 for CIFAR100.

C. Training Pipeline

- SimCLR Pretraining: The model was first trained using SimCLR for 30 epochs. To improve feature representation, the augmentations applied included RandomResizedCrop, ColorJitter, GaussianBlur, and RandomGrayscale.
- Warmup Phase: Following pretraining, a warmup phase of 5 epochs was conducted using Softmax-weighted Cross-entropy Loss (SWCE).
- Iterative Training with Stages:
  - For the first iteration, the stages were defined as epochs [2, 3, 4].
  - For subsequent iterations, stages were adjusted to [2, 5, 7].
  - The learning rate was 0.0001 for warmup and first iterations. Subsequently, the learning rate was decreased to 0.00001 at the second iteration and to 0.000001 at the last iteration.
  - A weight decay of 0.03 was used in the warmup phase to prevent model overfitting during warmup.
- Data Augmentation: During training, for each pseudo-labeled data point, two augmented versions were added to the batch to improve diversity and prevent overfitting.

D. Optimizer and Loss Function

- Optimizer: The Adam optimiser was used for most experiments with an initial learning rate 0.0001.
- Loss Function: SWCE Loss was utilised for main training and Normalized Cross-entropy + Mean Absolute Error was used for pseudo label assignment phase.

E. Evalation Metrics

We report the F1 score on test sets of all datasets as the primary performance metric.

Method	IDN - CIFAR10			IDN - CIFAR100		
	0.20	0.30	0.50	0.20	0.30	0.50
CE [9]	75.81	69.15	39.42	30.42	24.15	14.42
Mixup [10]	73.17	72.02	48.95	32.92	29.76	21.31
Forward [11]	74.64	69.75	46.27	36.38	33.17	19.27
Reweight [12]	76.23	70.12	45.46	36.73	31.91	20.23
Decoupling [13]	78.71	75.17	50.43	36.53	30.93	19.59
Co-teaching [14]	80.96	78.56	45.92	37.96	33.43	23.97
MentorNet [15]	81.03	77.22	47.89	38.91	34.23	24.15
Dividemix [8]	91.94	93.48	80.17	70.67	75.89	57.56
LSL [4]	97.13	96.85	95.81	80.94	79.90	77.95
Our Method	**94.40**	**93.60**	**82.11**	**76.61**	**73.11**	**61.33**

TABLE II: Performance (accuracy (%)) of different methods on IDN-CIFAR10 & IDN-CIFAR100 for different noise rates [4]

V. Results

The following results were obtained after implementing our pipeline on top of Dividemix. With the above experimental setup, the obtained results on CIFAR10 and CIFAR100 are shown in Table II. As seen from table II, we can successfully beat many previous methods.

We also tried experimenting with Resnet-34 on the CIFAR100 dataset. The results are given in Table III.

Method/Nr	Architecture	0.20	0.30	0.50
Dividemix [8]	Resnet-34	78.75	78.79	58.41
Our Method	Resnet-34	79.49	78.52	59.86
Our Method	Resnet-18	76.61	73.11	61.33

TABLE III: IDN – CIFAR100 results.

We also tried the pipeline on medical datasets like Chest X-ray and Chaoyang. The results are in Table IV.

Method	Chest X-ray		Chaoyang	
	0.20	0.30	0.20	0.30
Plain Training	67.58	64.82	53.12	46.88
Dividemix [8]	73.24	73.88	34.46	20.91
Our Method	**75.03**	**69.00**	**42.40**	**43.85**

TABLE IV: IDN Chest X-ray and IDN Chaoyang results.

We also tried using the Focal loss for medical datasets due to their imbalanced nature. Moreover, we tried replacing the active-passive loss in the pseudo label assignment by focal loss while keeping SWCE for the main training. The results are given in Table V.

Method	Chest X-ray		Chaoyang	
	0.20	0.30	0.20	0.30
Focal Loss	80.68	70.45	21.63	21.49
SWCE + FL	75.18	73.49	31.78	24.00

TABLE V: IDN Chest X-ray and IDN Chaoyang results.

VI. Key Observations

- Integrating contrastive learning during the initial phase helps establish a robust feature representation, mitigating the influence of noise in subsequent supervised tasks.
- Employing a multi-stage process enhances the reliability of pseudo-label selection, effectively reducing the propagation of noisy labels.
- Combining pseudo-labeled samples with their augmentations strengthens the model's learning of clean data patterns.
- We explored many techniques of Self-supervised Learning, such as BYOL and SimCLR, with SimCLR outperforming other techniques.
- The combined effect of SWCE and active-passive loss gave stable gradients during main training and identifying trustworthy labels during pseudo labels assignment.
- Our pipeline worked well for Chest X-ray, but is not performing well for Chaoyang. This is likely due to less training data, heterogeneity among samples and its imbalanced nature.
- Focal loss was giving better performance on Chest X-rays when used as a primary loss function for main training or even when used instead of active-passive loss. This is because Focal loss reduces the loss contribution from easy examples and focuses learning on hard, rare examples, making it ideal for imbalanced datasets.

VII. Future Works

- Our method can still not beat the State of the Art technique, which is [4]. We aim to make some constructive improvements in the pipeline, which can make it more robust.
- Devising a way to cluster the data samples based on their features and include that in our pipeline.
- Trying to improve the pipeline on medical datasets like HAM, Chaoyang, Blood Cell Images, etc.

VIII. Conclusion

This work addresses the critical problem of training neural networks with noisy labels, emphasising the importance of robust methods for real-world applications. Combining self-supervised learning with iterative pseudo-label refinement and utilising the proposed SWCE loss function demonstrates notable improvements in handling instance-dependent noise, as evidenced by its performance on benchmark datasets. It also enhances the model's noise tolerance and generalisation capabilities. While the results surpass many previous methods, further improvements are needed to rival the state-of-the-art techniques.

REFERENCES

[1] H. Song, M. Kim, D. Park, Y. Shin and J. -G. Lee, "Learning From Noisy Labels With Deep Neural Networks: A Survey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8135-8153, Nov. 2023.

[2] Ma, Xingjun, et al. "Normalized loss functions for deep learning with noisy labels." International conference on machine learning. PMLR, 2020.

[3] H. Song, M. Kim, and J.-G. Lee, "SELFIE: Refurbishing Unclean Samples for Robust Deep Learning," in Proceedings of the 36th International Conference on Machine Learning, pp. 5907-5915, PMLR, Jun. 2019.

[4] N.-R. Kim, J.-S. Lee, and J.-H. Lee, "Learning with Structural Labels for Learning with Noisy Labels," in Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 27600–27610.

[5] HChang, Haw-Shiuan, Erik Learned-Miller, and Andrew McCallum. "Active bias: Training more accurate neural networks by emphasising high variance samples." Advances in Neural Information Processing Systems 30 (2017).

[6] Wei, Hongxin, et al. "Open-set label noise can improve robustness against inherent label noise." Advances in Neural Information Processing Systems 34 (2021): 7978-7992.

[7] Barron, Jonathan T. "A general and adaptive robust loss function." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[8] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with Noisy Labels as Semi-supervised Learning," ArXiv, vol. abs/2002.07394, 2020.

[9] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. Advances in Neural Information Processing Systems, 34:4409–4420, 2021.

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimisation. arXiv preprint arXiv:1710.09412, 2017.

[11] 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 Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1944–1952, 2017.

[12] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence, 38(3):447–461, 2015.

[13] Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". Advances in neural information processing systems, 30, 2017.

[14] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co- teaching: Robust training of deep neural networks with extremely noisy labels. Advances in neural information processing systems, 31, 2018.

[15] 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 International conference on machine learning, pages 2304– 2313. PMLR, 2018.

[16] C. Ahn, K. Kim, J.-w. Baek, J. Lim, and S. Han, "Sample-wise label confidence incorporation for learning with noisy labels," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1823–1832, 2023.