PSEUDO-LABEL CORRECTION FOR INSTANCE-DEPENDENT NOISE USING TEACHER-STUDENT FRAMEWORK

Eugene Kim

University of California, San Diego

ABSTRACT

The high capacity of deep learning models to learn complex patterns poses a significant challenge when confronted with label noise. The inability to differentiate clean and noisy labels ultimately results in poor generalization. We approach this problem by reassigning the label for each image using a new teacher-student based framework termed P-LC (pseudo-label correction). Traditional teacher-student networks are composed of teacher and student classifiers for knowledge distillation. In our novel approach, we reconfigure the teacher network into a triple encoder, leveraging the triplet loss to establish a pseudo-label correction system. As the student generates pseudo labels for a set of given images, the teacher learns to choose between the initially assigned labels and the pseudo labels. Experiments on MNIST, Fashion-MNIST, and SVHN demonstrate P-LC's superior performance over existing state-of-the-art methods across all noise levels, most notably in high noise. In addition, we introduce a noise level estimation to help assess model performance and inform the need for additional data cleaning procedures.

***Index Terms*—** Weakly supervised learning, Instance-dependent noise, Label correction, Teacher-student framework

1. INTRODUCTION

As the size of training dataset increases, generalization of deep neural networks (DNNs) is expected to improve due to their efficient pattern memorization capabilities [1]. However, the ability to memorize complex patterns can lead to worse generalization performance depending on the accuracy of the dataset annotation. DNNs struggle to distinguish noise from clean data, which results in overfitting on noise and poor generalization [2, 3]. While accurately labeled data is essential, the majority of data label collection methods, e.g., crowdsourcing and web crawling, are expensive and susceptible to mistakes, particularly when dealing with large-scale datasets [4, 5, 6]. Hence, the research area of weakly-supervised learning holds much importance as it aims to improve model robustness in the presence of partially, imprecisely, or inaccurately labeled data [7]. In our paper, we tackle the challenge of learning with noisy labels.

To establish proper benchmarks, researchers have introduced various forms of controlled synthetic noise to replicate real-world noise conditions [8, 9, 10, 11]. The two most basic forms include symmetric noise and class-conditional noise (CCN). In a symmetric noise distribution, the corruption rate for all images are independently and identically distributed [8], whereas the corruption probability under a CCN assumption depends on the specific class of the image [11]. Nevertheless, in practice, the corruption probability for each image tends to differ regardless of its class association [12]. In order to more closely follow the real-world noise distribution, we adopt an approach involving convolutional neural networks (CNN) to generate instance-dependent noise (IDN) [10].

Earlier research in developing robust methods for noisy labels have leveraged statistical learning techniques, while more recent techniques have incorporated deep learning approaches. The statistical learning methods primarily encompass two forms: surrogate loss and noise rate estimation. For the surrogate loss approach, Masnadi-Shirazi et al. introduced SavageBoost, a boosting algorithm derived from a robust non-convex loss for corrupted binary classification [13]. Patrini et al. proposed a loss correction method using at most a matrix inversion and multiplication, but relied on the assumption that each class corruption probability is known [14]. For noise rate estimation methods, Menon et al. introduced a class-probability estimator by optimizing balanced error and AUC [15]. Both types of past statistical learning methods heavily relied on impractical noise assumptions, e.g. known class noise distribution, or were limited to binary classification, making them less effective for real-world scenarios.

More recent works have incorporated deep learning methods to correct or reweight weakly labeled data to achieve state-of-the-art (SOTA) classification accuracy [16, 17, 18]. Ren et al. used meta-learning to adjust the weights of training examples based on the gradient directions [19]. Li et al. trained the teacher classifier on a small and clean sample, while leveraging a Wikipedia-based knowledge graph to guide the training process of a student classifier [20]. To achieve competitive performance, this method required a carefully designed knowledge graph harvested from an external source of data. Our proposed method eliminates the need for additional data gathering and processing.

In contrast to the conventional teacher-student framework, our teacher network operates as a correction system for the predictions generated by the student network. We deem our method as pseudo-label correction (P-LC). P-LC can be divided into two phases: the teacher-student training phase and the noise correction phase. During the teacher-student training phase, both networks are separately trained on a clean dataset with their corresponding losses. During the correction phase, the student's task is to create reliable pseudo labels, while the teacher decides whether the initially assigned labels or the pseudo labels are more fit for the given images. The student is then retrained on the corrected dataset and makes predictions for the test set.

Our simple yet highly effective technique comes with two main advantages: reduced overall training time and self-adjustment to noise levels. The teacher and student network are trained concurrently on the clean dataset. Note the clean dataset is much smaller than the noisy dataset, also helping with the training time. Moreover, P-LC adjusts to noise levels without explicit tuning, allowing for consistent performance across all noise levels. As noise increases, the teacher network progressively places more trust in the pseudo la-

The author would like to thank Yifan Wu and Zhiting Hu from the University of California, San Diego for their helpful discussions and suggestions on this work.

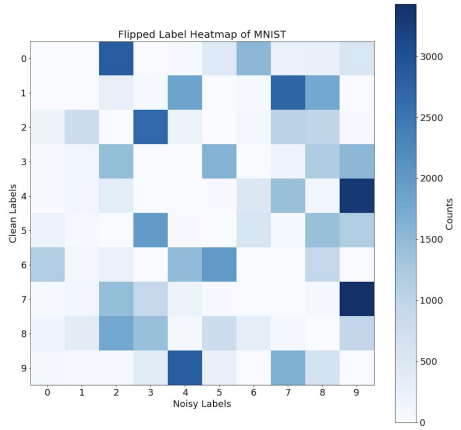

Fig. 1. Flipped label heatmap of MNIST under the IDN assumption.

bels over the initially assigned labels. Based on this model behavior, we further the use case by estimating the noise level of the dataset. We calculate the proportion of different labels between the corrected and noisy dataset as the estimation for the noise level.

The key findings and contributions of this paper can be outlined as follows:

- We introduce a novel teacher-student framework that leverages pseudo labels to correct potentially misleading noisy labels.
- To the best of our knowledge, we propose the first teacher-student based approach for noise level estimation.
- We conduct experiments on three common benchmark datasets with varying noise levels and the proposed method outperforms SOTA methods, most visibly in high noise levels.

2. PROBLEM FORMULATION AND ASSUMPTIONS

2.1. Preliminary

For a K-class classification problem, we define the feature space X and label space $Y = \{1, \ldots, k\}$. Consistent with prior research, we are given a noisy training dataset $D_{noise} = \{(x_i, y_i')\}_{i=1}^{n}$ and clean dataset $D_{clean} = \{(x_i, y_i)\}_{i=1}^{m}$, where $m << n$ (m is the size of D_{clean} and n is the size of D_{noise}) [21, 22, 16, 20]. The corrected dataset after implementing our method is defined as $D_{corrected} = \{(x_i, \bar{y}_i)\}_{i=1}^{n}$. Note D_{clean} has a uniform label distribution, and therefore, does not guarantee the same label distribution as D_{noise} or $D_{corrected}$.

2.2. Noise Assumptions

IDN with varying noise levels are injected into clean datasets to closely imitate real-world noise generation while retraining a controlled environment. We adopt an IDN generation approach that leverages the soft label predictions from a CNN trained on the entire clean dataset [10]. For each input, the CNN's prediction with the highest likelihood is used as the noisy label. By enforcing a likelihood for each image, it goes beyond the class-conditional assumptions and satisfies the requirements for IDN.

The distribution of flipped labels based on the IDN assumption indicates that for most of classes, there exists a single corresponding class with the highest flip likelihood. Thus, the majority of potential corruption classes do not provide any additional information. As seen in Fig. 1, images depicting the number 4 are predominantly flipped to resemble the number 9, while images of the number 1 tend to be transformed into the number 4, and so forth. This observation forms the basis for choosing a hard over soft label correction method. Moreover, the performance of existing soft label correction methods drops significantly as the noise level increases [10, 16]. Recent study by Wei et al. reports that label smoothing (LS) actually decreases model performance in high noise settings [23]. The computational cost of the soft label is also more expensive compare to the one-hot encoded label. Thus, we develop a simple yet highly effective approach by employing a hard pseudo-label correction technique.

3. METHODOLOGY

3.1. Overview

Our proposed method P-LC involves a teacher-student network and can be divided into two separate phases: the teacher-student training phase and noise correction phase. During the teacher-student training phase, the teacher network learns to differentiate images from the same class (positive pairs) and images from different classes (negative pairs). The student network focuses on standard image classification. Both networks are trained on the small set of clean data D_{clean}. During the noise correction phase, the student network generates pseudo labels for all images in the noisy dataset D_{noise}. The teacher network then decides whether to use the pseudo labels or initially assigned labels as the new labels in the corrected dataset $D_{corrected}$. As a final step, the student network is retrained on $D_{corrected}$ to provide predictions for the test set.

3.2. Teacher-Student Architecture

For the teacher network T, we use a Siamese framework composed of three identical encoders, all consisting of the same architecture and weights [24]. When training on the clean data D_{clean}, the inputs consist of an anchor image x_{anc}, a positive image x_{pos}, and a negative image x_{neg} defined as follows

$$x_{anc} \in X^k, x_{pos} \in X^k \setminus \{x_{anc}\}, x_{neg} \in X \setminus X^k \tag{1}$$

where X^k refers to all images in feature space X with class label k. This ensures the anchor image is from the same class as the positive image but different class as the negative image. Once x_{anc} has a corresponding x_{pos} and x_{neg}, the network encodes x_{anc}, x_{pos}, and x_{neg} as z_{anc} (anchor embedding), z_{pos} (positive embedding), and z_{neg} (negative embedding), respectively. For every anchor embedding, we want to decrease the distance from the positive embedding but increase the distance from the negative embedding as shown below

$$\|z_{anc} - z_{pos}\|_2^2 + \alpha < \|z_{anc} - z_{neg}\|_2^2 \tag{2}$$

where α is a hyperparameter that sets the minimum difference between the positive and negative embeddings. The teacher loss ℓ_T is then minimized, which is the triplet loss displayed below

$$\ell_T = \sum_{i=1}^{m} \max(\|z_{anc}^i - z_{pos}^i\|_2^2 - \|z_{anc}^i - z_{neg}^i\|_2^2 + \alpha, 0) \tag{3}$$

For the student network S, we use a combination of CNN layers to learn lower level image embeddings and fully-connected layers to

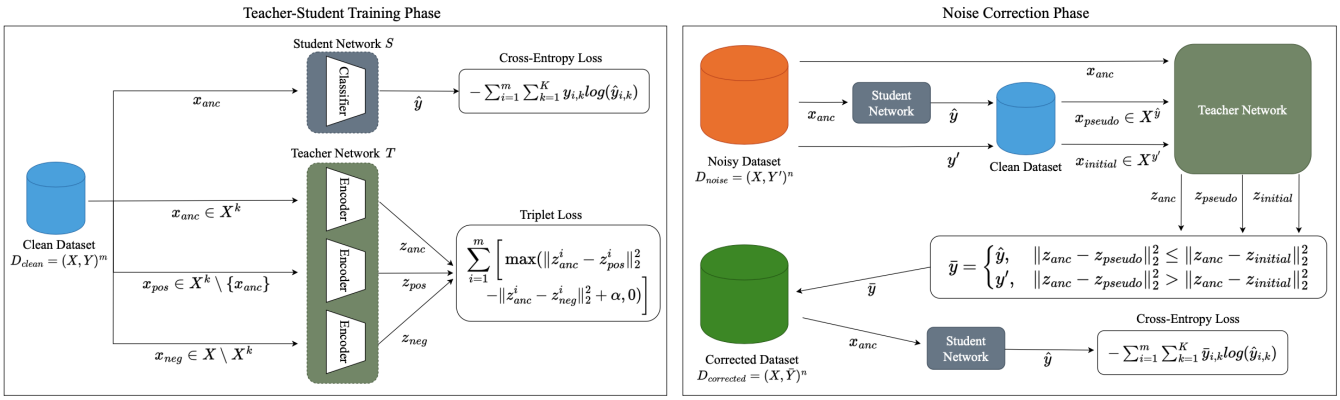

Fig. 2. P-LC's computation graph. The left panel shows the training process of the teacher-student network on the clean dataset. On the right panel, we show the noise correction process composed of pseudo-label generation, data sampling, and label correction.

make label predictions. We minimize the student loss ℓ_S, which is the cross-entropy loss displayed below

$$\ell_S = -\sum_{i=1}^{m}\sum_{k=1}^{K} y_{i,k} log(\hat{y}_{i,k}) \tag{4}$$

where $\hat{y}$ is the model prediction and y is the true label. The teacher and student network can be trained concurrently on D_{clean} because ℓ_S and ℓ_T rely on different inputs.

3.3. Learning to Correct Pseudo-labels

Once both the teacher and student network are trained on D_{clean}, the noise correction phase begins. For the entire correction phase, every image in D_{noise} is treated as an anchor image x_{anc}. The correction phase is composed of three main components: pseudo label generation, data sampling, and label correction.

For each input x_{anc} in D_{noise}, the student network generates a pseudo label $\hat{y}$. $\hat{y}$, alongside the initial label y', present two potential classes of x_{anc}. In the case that $\hat{y} = y'$, no correction is required by the teacher network. x_{anc} retains y' as the label in the corrected dataset $D_{corrected}$. Otherwise, we randomly sample two images x_{pseudo} and $x_{initial}$ from D_{clean} with different class labels (i.e. $x_{pseudo} \in X^{\hat{y}}$ and $x_{initial} \in X^{y'}$). The teacher network then encodes x_{anc}, x_{pseudo}, and $x_{initial}$ as z_{anc}, z_{pseudo}, and $z_{initial}$. Based on the dissimilarity correction metric, the corrected label $\bar{y}$ is computed as

$$\bar{y} = \begin{cases} \hat{y}, & \|z_{anc} - z_{pseudo}\|_2^2 \leq \|z_{anc} - z_{initial}\|_2^2 \\ y', & \|z_{anc} - z_{pseudo}\|_2^2 > \|z_{anc} - z_{initial}\|_2^2. \end{cases} \tag{5}$$

(5) indicates the teacher network assigns the label of the image that most closely resembles the anchor image. $(x_{anc}, \bar{y})$ is added to $D_{corrected}$. The student network retrains on $D_{corrected}$ and is evaluated on the test data. The complete process is shown in Fig. 2.

We can improve the label correction accuracy by increasing the number of samples from D_{clean}. With a larger pool of images for comparison alongside x_{anc}, we reduce the variance and improve generalization as seen in previous studies [25, 26]. In practice, we sample five images instead of one image from D_{clean}. We then assign x_{anc} to the label of the images with the highest number of $\bar{y}$ transformations. This approach is particularly effective when dealing with low-quality or corrupted sampled images.

P-LC has two main advantages over existing methods: reduced overall training time and self-adjustment to noise levels. First, the separate loss functions used in the teacher-student network enables concurrent training. Second, P-LC does not require label smoothing, reducing the computation cost when calculating ℓ_S. Beyond reduced training time, P-LC can adjust to varying noise levels with a properly trained teacher-student network. In high noise levels, more label corrections are made by the teacher network due to the increased cases where $\hat{y} \neq y'$. This dynamic adjustment of label corrections based on noise level represents an internal tuning mechanism inherent to P-LC.

3.4. Noise Level Estimation

We further our work by introducing a noise level estimator built upon the same teacher-student architecture. Once we obtain $D_{corrected}$ using P-LC, we can estimate the noise level with one additional calculation. We simply compute the proportion p_{noise} of different labels between $D_{corrected}$ and D_{noisy} as displayed below:

$$p_{noise} = \frac{\sum_{i=1}^{n} I(\bar{y}_i, y'_i)}{n}, \tag{6}$$

where $I(\bar{y}_i, y'_i) = \begin{cases} 1, & \bar{y}_i \neq y'_i \\ 0, & \text{otherwise.} \end{cases}$ Computing p_{noise} provides benefits when the true noise level is both unknown and known. In the case of an unknown true noise level, an estimation can help assess model performance and inform the need for additional data cleaning procedures. In our study, we leverage p_{noise} as an initial indicator to evaluate P-LC's relabeling accuracy. Depending on the proximity of p_{noise} to the true noise level, we make adjustments to the hyperparameters of the teacher-student network. Table 2 presents the averaged noise level estimations across three datasets. Although P-LC struggles to estimate exact noise levels, it correctly ranks the noise level ranging from 20% IDN to 50% IDN.

4. EXPERIMENTS

4.1. Datasets

We evaluate P-LC against SOTA methods [27, 28, 10] using three image recognition datasets: MNIST, Fashion-MNIST, and SVHN. For existing methods that do not require a clean dataset, we inject

Table 1. Classification accuracy on MNIST, Fashion-MNIST, and SVHN with different instance-dependent label noise levels.

Method	MNIST				Fashion-MNIST				SVHN			
	IDN-20%	IDN-30%	IDN-40%	IDN-50%	IDN-20%	IDN-30%	IDN-40%	IDN-50%	IDN-20%	IDN-30%	IDN-40%	IDN-50%
SOTA	93.07	88.53	77.48	73.27	82.84	82.15	79.98	72.96	83.09	78.73	73.64	64.90
	± 0.22	± 0.25	± 0.79	± 0.05	± 0.12	± 0.01	± 0.19	± 0.37	± 0.27	± 0.74	± 0.54	± 0.53
P-LC	**94.81**	**94.26**	**93.71**	**93.07**	**83.60**	**82.61**	**80.89**	**79.78**	**83.31**	**82.03**	**78.62**	**76.15**
	± 0.20	± 0.15	± 0.25	± 0.21	± 0.06	± 0.12	± 0.14	± 0.42	± 0.13	± 0.13	± 0.37	± 0.11

Table 2. P-LC noise level estimations on MNIST, Fashion-MNIST, and SVHN

	IDN-20%	IDN-30%	IDN-40%	IDN-50%
MNIST	13.40 ± 0.16	22.59 ± 0.19	31.92 ± 0.24	41.69 ± 0.28
F-MNIST	8.09 ± 0.18	12.15 ± 0.25	17.96 ± 0.18	25.06 ± 0.41
SVHN	15.37 ± 0.06	22.35 ± 0.10	23.56 ± 0.24	30.21 ± 0.21

the training data with IDN at rates of 20%, 30%, 40%, and 50%. In contrast, for our method, we first construct the clean dataset D_{clean} by sampling from the training set. The remaining training set is injected with four levels of IDN at rates of 22%, 32%, 42%, and 52% to compensate existing methods that do not require a clean dataset. We intentionally place our method at a disadvantage by slightly increasing the noise across all experiments. A summary of each dataset is provided in Table 3.

Table 3. Overview of our datasets and teacher-student network architectures. Note the preceding numbers in CNN-4 and Siamese-5 refer to the total number of layers.

Dataset	MNIST	Fashion-MNIST	SVHN
# classes	10	10	10
RGB	No	No	Yes
Train	60,000	60,000	73,257
Test	10,000	10,000	26,032
Clean	1,200	1,200	1,465
Noisy	58,800	58,800	71,792
Teacher net.	Siamese-5	Siamese-5	Siamese-5
Student net.	CNN-4	CNN-4	ResNet-18

4.2. Implementation Details

We compare our method against four existing methods: cross-entropy (CE) loss, co-teaching [27], deep abstaining classifier (DAC) [28], and self-evolution average label (SEAL) [10]. To ensure a fair and consistent comparison, all implementations are done on PyTorch and experiments are executed on NVIDIA GeForce GTX 1080 Ti. We use the same classifier architecture across all methods, i.e., ResNet-18 for SVHN and 4-layered CNN for MNIST and Fashion-MNIST. Existing methods are trained three times (seed 0, 1, and 2) for 50 epochs on every noise level. For our method, we use 50 epochs for retraining on the corrected dataset $D_{corrected}$. The batch size is set to 64. Due to limited space, we report the highest accuracy achieved among the four existing methods as SOTA in Table 1.

4.3. Results

In Table 1, we present the averaged accuracies across 4 distinct IDN levels, ranging from 20% to 50%. Based on our experimental results, P-LC consistently outperforms existing methods across all three datasets. The most notable differences are seen in high noise settings. While SOTA methods experience a significant decline in accuracy as the noise level increases, P-LC's test accuracy depreciates at a lower rate.

Several factors contribute to the reduced performance observed in SOTA methods at high noise levels. In the case of co-teaching, it relies on sampling clean images from noisy data to guide student training [27]. However, as noise increases, the pool of available clean images for selection diminishes. DAC is a different approach that abstains from making predictions when the class label's uncertainty is high [28]. Nonetheless, DAC faces the same limitations as co-teaching. High noise levels restrict the available information from the training data, resulting in poor generalization. We address this problem by employing a relabeling over reweighting technique. This enables the model to undergo training on a complete dataset, preventing the limitation of data as noise increases. SEAL, the third SOTA method, uses a label smoothing technique to iteratively readjust the label distribution of images [10]. According to research conducted by Wei et al., label smoothing leads to reduced accuracy in high noise settings. [23]. We overcome this issue by generating hard instead of soft-pseudo labels. By incorporating both hard-pseudo labels and relabeling techniques into P-LC, the proposed method outperforms SOTA methods across all noise levels.

5. CONCLUSION

In this paper, we address the issue of noisy labels from a label correction perspective. We explore the limitations of existing methods, specifically label smoothing and reweighting techniques, in the presence of instance-dependent noise. We then introduce a novel teacher-student framework designed to address these challenges by integrating hard-pseudo labels with label corrections. Essentially, the teacher encoder operates as a correction system for predictions made by the student classifier. P-LC comes with two main advantages: reduced overall training time and self-adjustment to noise levels. Empirical experiments on MNIST, Fashion-MNIST, and SVHN with varying noise levels demonstrate the superior noise robustness of P-LC compared to SOTA methods, particularly in high noise environments. Furthermore, we introduce a simple noise level estimation to help assess model performance and inform the need for additional data cleaning procedures.

6. REFERENCES

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[2] Benoît Frénay and Michel Verleysen, "Classification in the presence of label noise: a survey," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 5, pp. 845–869, 2013.

[3] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus, "Training convolutional networks with noisy labels," in *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[4] Peter Welinder and Pietro Perona, "Online crowdsourcing: rating annotators and obtaining cost-effective labels," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 25–32.

[5] Sriram Raghavan and Hector Garcia-Molina, "Crawling the hidden web," in *Vldb*, 2001, vol. 1, pp. 129–138.

[6] Rob Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman, "Learning object categories from internet image searches," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1453–1466, 2010.

[7] Zhi-Hua Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[8] Dana Angluin and Philip Laird, "Learning from noisy examples," *Machine learning*, vol. 2, pp. 343–370, 1988.

[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," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7597–7610, 2020.

[10] Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng, "Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 11442–11450.

[11] Clayton Scott, Gilles Blanchard, and Gregory Handy, "Classification with asymmetric label noise: Consistency and maximal denoising," in *Conference on learning theory*. PMLR, 2013, pp. 489–511.

[12] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *International conference on machine learning*. PMLR, 2020, pp. 4804–4815.

[13] Hamed Masnadi-Shirazi and Nuno Vasconcelos, "On the design of loss functions for classification: theory, robustness to outliers, and savageboost," *Advances in neural information processing systems*, vol. 21, 2008.

[14] 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*, 2017, pp. 1944–1952.

[15] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson, "Learning from corrupted binary labels via class-probability estimation," in *International conference on machine learning*. PMLR, 2015, pp. 125–134.

[16] Guoqing Zheng, Ahmed Hassan Awadallah, and Susan Dumais, "Meta label correction for learning with weak supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[17] Tongliang Liu and Dacheng Tao, "Classification with noisy labels by importance reweighting," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 447–461, 2015.

[18] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," *Advances in neural information processing systems*, vol. 32, 2019.

[19] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun, "Learning to reweight examples for robust deep learning," in *International conference on machine learning*. PMLR, 2018, pp. 4334–4343.

[20] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1910–1918.

[21] Moses Charikar, Jacob Steinhardt, and Gregory Valiant, "Learning from untrusted data," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, 2017, pp. 47–60.

[22] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 839–847.

[23] Jiaheng Wei, Hangyu Liu, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Yang Liu, "To smooth or not? when label smoothing meets noisy labels," *arXiv preprint arXiv:2106.04149*, 2021.

[24] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[25] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[26] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *International conference on learning representations*, 2016.

[27] 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*, vol. 31, 2018.

[28] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof, "Combating label noise in deep learning using abstention," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6234–6243.