Learning with Noisy Labels by Efficient Transition Matrix Estimation to Combat Label Miscorrection

Seong Min Kye†, Kwanghee Choi†, Joonyoung Yi†, and Buru Chang*

Hyperconnect Inc.
{harris,kwanghee.choi,joonyoung.yi,buru.chang}@hpcnt.com

Abstract. Recent studies on learning with noisy labels have shown remarkable performance by exploiting a small clean dataset. In particular, model agnostic meta-learning-based label correction methods further improve performance by correcting noisy labels on the fly. However, there is no safeguard on the label miscorrection, resulting in unavoidable performance degradation. Moreover, every training step requires at least three back-propagations, significantly slowing down the training speed. To mitigate these issues, we propose a robust and efficient method, *FasTEN*, which learns a label transition matrix on the fly. Employing the transition matrix makes the classifier skeptical about all the corrected samples, which alleviates the miscorrection issue. We also introduce a two-head architecture to efficiently estimate the label transition matrix every iteration within a single back-propagation, so that the estimated matrix closely follows the shifting noise distribution induced by label correction. Extensive experiments demonstrate that our FasTEN shows the best performance in training efficiency while having comparable or better accuracy than existing methods, especially achieving state-of-the-art performance in a real-world noisy dataset, Clothing1M.

Keywords: Learning with noisy labels; Label correction; Transition matrix estimation

1 Introduction

In the last decade, supervised learning has achieved great success by leveraging an abundant amount of annotated data to solve various classification tasks such as image classification [37], object detection [24], and face recognition [89]. It has been proven both theoretically and empirically that the performance of supervised learning-based classification models steadily improves as the size of annotated data increases [27, 12, 21]. However, we cannot avoid *noisy labels* due to its coarse-grained annotation sources [40, 121], resulting in performance degradation [14].

† Equal contribution.

* Corresponding author.

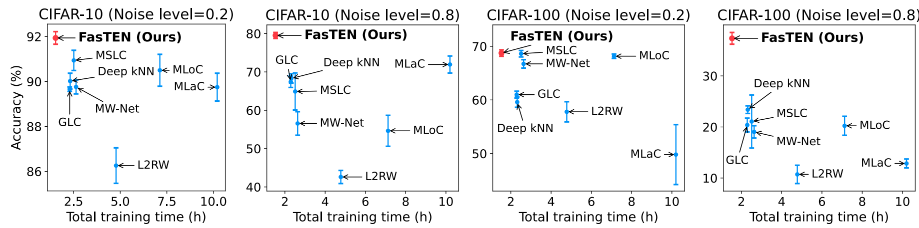


Fig. 1: Plotting accuracy (%) **(y-axis)** according to total training time (hours) **(x-axis)**. Our proposed method (FasTEN) shows the best performance in training efficiency while having comparable or better accuracy on both CIFAR-10/100 with various noise levels.

Many methods have been proposed to build a classifier that is robust to noisy labels. Unlike traditional methods [68, 94, 2, 75] which assume that all the given labels are potentially corrupted, recently proposed methods utilize an inexpensively obtained small clean dataset to improve performance further. Based on the clean data set, loss correction methods [40, 101] reduce the influence of noisy labels by modifying loss functions and re-weighting methods [80, 85, 3, 23] penalize samples that are likely to be noisy labels. Especially, recent label correction methods [105, 121] achieve remarkable performance based on model-agnostic meta-learning (MAML) [20]. These methods relabel noisy labels to directly reduce the noise level, raising the theoretical upper bound of the predictive performance (See Appendix A.1).

However, there are two challenges for these MAML-based label correction methods: (1) *The label correction methods blindly trust the already miscorrected labels.* Erroneously corrected labels are often kept throughout the training, which causes the model to learn the miscorrected labels as ground-truth labels. Several studies [67, 105] attempt to tackle this through training techniques such as soft labels, whereas it does not fundamentally solve the problem. (2) *MAML-based methods are inherently slow in training, resulting in excessive computational overhead.* The inefficiency comes from multiple training steps per single iteration of MAML-based methods, including virtual updates with inner optimization loops.

To alleviate these issues, we propose a robust and efficient method called ***FasTEN*** (**Fas**t **T**ransition Matrix **E**stimation for Learning with **N**oisy Labels). FasTEN efficiently estimates a transition matrix to learn with noisy labels while continuously correcting them on-the-fly. It is theoretically proven that the correctly estimated label transition matrix is useful to obtain a statistically consistent classifier from noisy labels [108, 113] (See Appendix A.2), i.e., more robust to noisy labels. To efficiently estimate the transition matrix, we adopt a two-head architecture that consists of two classifiers, a noisy and a clean classifier, with a shared feature extractor. For every iteration, the noisy classifier estimates the label transition matrix shifted by the label correction. On the other hand, the clean classifier is trained to be statistically consistent by leveraging the estimated transition matrix. Using the output of the clean classifier, FasTEN relabels noisy labels to reduce the noise level. Our proposed FasTEN has a safeguard for the miscorrected labels since it adaptively estimates the transition matrix on every iteration, so that the clean classifier stays equally skeptical towards all the corrected labels. Furthermore, our efficient method jointly optimizes the two-head architecture with only a single back-propagation for each iteration, boosting training speed. In this paper, we focus on solving the problem of *class-dependent* noisy labels [25, 76, 119] (i.e., $p(\bar{y}|y,x) = p(\bar{y}|y)$), although the problem of instance-dependent noisy labels [107, 16, 123] remains an important problem to be addressed.

Experimental results show that our method achieves state-of-the-art performance by a large margin on both the synthetic and real-world noisy label datasets, various noise levels of *CIFAR* [49] and *Clothing1M* [109], respectively. We demonstrate the exceptional training speed of our proposed FasTEN while achieving better performance compared to baselines, as shown in Figure 1. Especially, although our FasTEN assumes only class-dependent noisy labels, it also achieves state-of-the-art performance in the Clothing 1M dataset which contains instance-dependent noisy labels. This experimental result supports recent observations that leveraging the accurately estimated transition matrix with small clean data is helpful for alleviating instance-dependent noise [66, 40, 123, 45] (See Appendix A.2). Finally, we conduct a thorough analysis to understand the inner mechanisms of our proposed method.

Our contribution in this paper is threefold: (1) We propose a robust and efficient method that learns a transition matrix to learn with noisy labels while continuously correcting them on the fly. To the best of our knowledge, this is the first attempt to improve the label correction with the transition matrix estimation. (2) Our proposed method boosts training speed by employing a two-head architecture so that the label transition matrix can be learned with a single back-propagation. (3) Extensive experiments validate the efficacy of our proposed method in terms of both training speed and predictive performance.

2 Related Work

Learning with noisy labels assumes that labels in all the training samples are potentially corrupted. They can be further categorized as follows: *various loss functions* [71, 94, 75, 22, 120, 99, 94, 62, 63, 61, 57, 112, 45], *regularizations* [2, 46, 38, 42, 65, 62, 35, 87, 59, 57, 10, 39, 73, 74, 53, 117, 36, 55, 64], *re-weighting training samples* [80, 44, 67, 60, 100, 93, 14, 43, 102, 104, 104, 77], and *correcting noisy labels* [90, 114, 34, 86, 122, 30, 47, 106, 119]. However, different losses or regularizations yield inferior performance to state-of-the-art methods [117, 67, 42, 55], and re-weighting methods often filter out noisy but helpful samples for extracting features to show sub-optimal performance [86, 105, 121, 67, 11, 58, 84]. *Label correction* methods circumvent their shortcomings by relabeling so that the feature extractor leverage the corrected labels. However, label correction methods also have a limitation in that they are prone to propagate the error when miscorrected labels are continuously accumulated [67, 105, 121]. Others *correct the training loss* by estimating a label transition matrix [68, 79, 88, 7, 76, 25, 111, 108, 113] to build a statistically consistent classifier, where the methods need multiple training stages; e.g., include a separate pretraining stage. In this paper, we join a simple label correction method with estimating the label transition matrix to alleviate the miscorrection issue caused by miscorrected noisy labels, which only requires a single training stage.

Learning with Noisy Label via Small Clean Dataset. Unlike traditional methods that use noisy datasets only [76, 113, 56], several recent studies argue that a small clean dataset is easily obtained by techniques such as image retrieval [78]; hence one can further devise a method that effectively leverages it. Many studies have successfully adapted the idea and shown massive performance improvement compared to the traditional methods. Early methods [40, 3, 118] require multiple training stages where it hinders the training efficiency. Recent studies widely adopt MAML [20] to various strategies discussed above: sample re-weighting [97, 50, 44, 80, 54, 85], label correction [105, 121], and label transition matrix estimation [101]. These approaches first perform a virtual update with the noisy dataset, find optimal parameters using the clean dataset, and update the actual parameters by the found parameters. This virtual update process requires three back-propagations per iteration, leading to at least three times the computational cost. Our proposed label correction method estimates the label transition matrix using a batch drawn from the clean small dataset in a single back-propagation, greatly enhancing the training speed while showing comparable or better performance to existing state-of-the-art methods. Additional related works are described in Appendix D.

3 Methodology

Existing label correction methods try to find and fix noisy labels to utilize them as clean samples in model training, where they can improve the classification performance by reducing the noise level of the whole training samples. However, erroneously corrected samples, i.e., clean samples deemed noisy, or vice versa, are often kept throughout the model training. Since current label correction methods blindly trust these miscorrected labels, this behavior degrades the classification performance under the noisy label situation (§ 4.4).

In this section, we show that the accurately estimated label transition matrix with the clean dataset alleviates the miscorrection problem of existing label correction methods. Further, we describe our efficient method estimating the label transition matrix for every training iteration while correcting noise labels. Our proposed method is illustrated in Figure 2 and summarized in Algorithm 1.

3.1 Batch Formation

We estimate the transition matrix to track the shifted noisy label distribution caused by label correction using a clean batch. To ensure the effective estimation

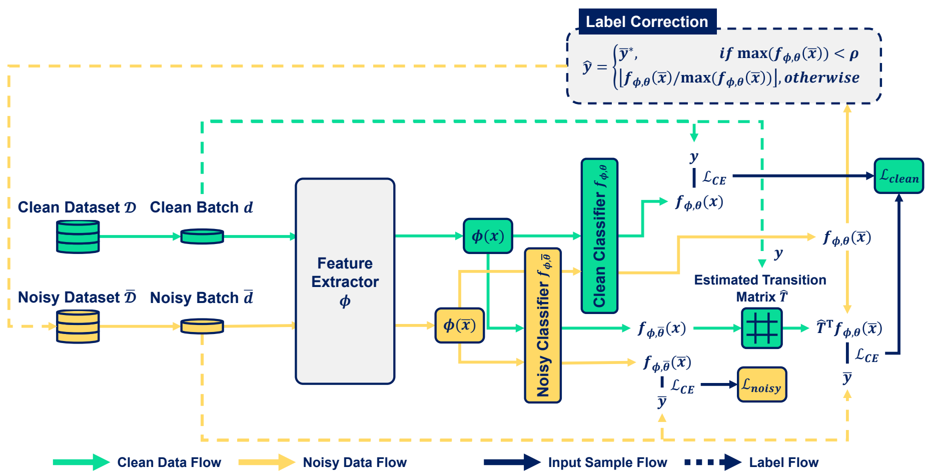


Fig. 2: Summarization of our proposed method (FasTEN).

of the label transition matrix, we formulate the batch to have the same number of samples per class. We first compose the clean batch d with randomly chosen K samples for the entire N classes in the clean dataset $\mathcal{D}$ to get benefits from having a certain amount of clean samples for each class, as follows $d = \{(x_n, y_n)\}_{n=1}^{KN}$ where x is an input and $y \in \mathbb{R}^N$ is the clean label of x. A noisy batch $\bar{d}$ is randomly sampled from the noisy dataset $\bar{\mathcal{D}}$, as follows $\bar{d} = \{(\bar{x}_n, \bar{y}_n)\}_{n=1}^{M}$ where $\bar{y} \in \mathbb{R}^N$ is the noisy label of $\bar{x}$ and M is the size of the noisy batch which we set as $M = KN$ for simplicity. It is different from other methods [121, 105, 85, 80] based on meta-learning which randomly compose the clean batch.

3.2 Transition Matrix Estimation

Each element T_{ij} of the label transition matrix $\boldsymbol{T} \in \mathbb{R}^{N \times N}$ is defined as the probability of a clean label i to be corrupted as a noisy label j, i.e. $T_{ij} = p(\bar{y} = j|y = i)$. It is well-known that a robust classifier can be obtained with the accurately estimated label transition matrix [88, 76, 40, 108, 113]. We choose a simple but accurate method that directly estimates the posterior with a clean dataset [40, 108, 113], whereas there are other more sophisticated methods that estimate the label transition matrix [68, 79, 88, 7, 25, 76]. Following the assumption of the previous work [40, 15, 8, 107], we also assume conditional independence of $\bar{y}$ and y given x:

$$p(\bar{y}|y) = p(\bar{y}|y) \int p(x|\bar{y}, y)dx = \int p(\bar{y}|y, x)p(x|y)dx = \int p(\bar{y}|x)p(x|y)dx. \quad (1)$$

We design the transition matrix to be *class-dependent*, i.e., $p(\bar{y}|y, x) = p(\bar{y}|y)$, following recent state-of-the-art methods [60, 82, 108, 113]. By parameterizing a feature extractor $\bar{\phi}$ and a linear classifier $\bar{\theta}$, we obtain $p(\bar{y}|x) = f_{\bar{\phi},\bar{\theta}}(x)$ where $f_{\bar{\phi},\bar{\theta}}$

Algorithm 1 **Fas**t **T**ransition Matrix **E**stimation for Learning with **N**oisy Labels (**FasTEN**)

Input: Clean dataset $\mathcal{D}$, noisy dataset $\bar{\mathcal{D}}$.
Hyper-parameters: Label correction threshold ρ, Controllable loss ratio for noisy classifier λ.
Output: Clean classifier $f_{\phi,\theta}$ where linear classifier θ and feature extractor ϕ.
Randomly initialize common feature extractor ϕ, linear classifiers θ, and $\bar{\theta}$ for clean labels and noisy labels.
for each epoch $i = 0, \cdots$ **do**
  for each iteration in epoch i **do**
    Sample mini-batch $d \sim \mathcal{D}$, $\bar{d} \sim \bar{\mathcal{D}}$.
    $\hat{T} \leftarrow \left(\sum_{(x,y)\in d} y f_{\phi,\bar{\theta}}(x)^\top\right) \mathrm{diag}^{-1}\left(\sum_{(x,y)\in d} y\right)$
    $\mathcal{L}_{\text{clean}} \leftarrow \sum_{(x,y)\in d} \mathcal{L}_{\text{CE}}\left(f_{\phi,\theta}(x), y\right) + \sum_{(\bar{x},\bar{y})\in \bar{d}} \mathcal{L}_{\text{CE}}\left(\hat{T}^\top f_{\phi,\theta}(\bar{x}), \bar{y}\right)$
    $\mathcal{L}_{\text{noisy}} \leftarrow \sum_{(\bar{x},\bar{y})\in \bar{d}} \mathcal{L}_{\text{CE}}\left(f_{\phi,\bar{\theta}}(\bar{x}), \bar{y}\right))$
    $\bar{\mathcal{D}} \leftarrow (\bar{\mathcal{D}} - \bar{d}) \cup \left\{ \left(\bar{x}, \begin{cases} \bar{y}^*, & \text{if } \max(f_{\phi,\theta}(\bar{x})) < \rho \\ \lfloor f_{\phi,\theta}(\bar{x}) / \max(f_{\phi,\theta}(\bar{x})) \rfloor, & \text{otherwise} \end{cases} \right) \middle| (\bar{x}, \bar{y}) \in \bar{d} \right\}$
    Update $\phi, \theta, \bar{\theta}$ using $\nabla_{\phi,\theta,\bar{\theta}}\left(\mathcal{L}_{\text{clean}} + \lambda \mathcal{L}_{\text{noisy}}\right)$ with a single back-propagation.
  end for
end for

is the noisy classifier that consists of the linear classifier and the feature extractor trained only with the noisy labels. If the noisy classifier $f_{\bar{\phi},\bar{\theta}}$ gives a perfect prediction for the noisy data, we can estimate the transition probability $p(\bar{y}|y)$ using the clean samples $(x, y) \in d$ as follows (See Appendix A.3 for details):

$$\widehat{\boldsymbol{T}} \leftarrow (\sum_{(x,y)\in d} y f_{\bar{\phi},\bar{\theta}}(x)^\top) \mathrm{diag}^{-1}(\sum_{(x,y)\in d} y). \tag{2}$$

We emphasize the importance of the transition matrix estimation, as its accuracy determines the bounds of the generalization error of the classifier [108]. However, the limited number of clean samples inside a single batch may yield an inaccurate transition matrix, even with the ideal $f_{\bar{\phi},\bar{\theta}}$. We analyze the upper bound of the estimation error as follows:

Theorem 1. *Assume the Frobenius norm of the weight matrices $\bar{\phi}_1, ..., \bar{\phi}_{H-1}, \bar{\theta}$ are at most $\bar{\Phi}_1, ..., \bar{\Phi}_{H-1}, \bar{\Theta}$ for H-layer neural networks $f_{\bar{\phi},\bar{\theta}}$. Let the loss function be L-Lipschitz continuous w.r.t. $f_{\bar{\phi},\bar{\theta}}$. Let the activation functions be 1-Lipschitz, positive-homogeneous, and applied element-wise (such as ReLU). Let x be upper bounded by B, i.e., for any $x \in \mathcal{X}$, $\|x\| \leq B$. Then, for $\epsilon \geq 0$*

$$\begin{aligned} p\left(\left|\widehat{T}_{ij} - T_{ij}\right| > \epsilon\right) \leq & \frac{NLB(\sqrt{2H\log 2} + 1)\bar{\Theta}\Pi_{h=1}^{H-1}\bar{\Phi}_i}{\sqrt{|\bar{\mathcal{D}}|}} \\ & + \frac{\sqrt{-\log(\epsilon)}}{\sqrt{2|\bar{\mathcal{D}}|}} + 2\exp\left(-2\epsilon^2 K\right). \end{aligned} \tag{3}$$

Proof. See Appendix A.4. □

Although the upper bound of the estimation error of the transition matrix is affected by the batch size K, we empirically verify that small K does not necessarily harm the classification performance (See Appendix C.7).

3.3 Learning with Estimated Transition Matrix

A clean classifier $f_{\phi,\theta}$ is trained with the estimated transition matrix $\widehat{\boldsymbol{T}}$:

$$\mathcal{L}_{\text{clean}} = \sum_{(x,y)\in d} \mathcal{L}_{\text{CE}}\left(f_{\phi,\theta}(x), y\right) + \sum_{(\bar{x},\bar{y})\in \bar{d}} \mathcal{L}_{\text{CE}}\left(\widehat{\boldsymbol{T}}^{\top} f_{\phi,\theta}(\bar{x}), \bar{y}\right), \tag{4}$$

given the cross-entropy loss function $\mathcal{L}_{\text{CE}}$, where the feature extractor ϕ and the linear classifier θ form the clean classifier $f_{\phi,\theta}$ which estimates clean labels. If $\widehat{\boldsymbol{T}}$ is correctly estimated, the clean classifier $f_{\phi,\theta}$ becomes statistically consistent [88, 76, 40, 108, 113]. This approach makes the clean classifier skeptical towards corrected labels, hence avoiding the miscorrection issue.

On the other hand, the noisy classifier $f_{\bar{\phi},\bar{\theta}}$ is trained to model the noisy label distribution.

$$\mathcal{L}_{\text{noisy}} = \sum_{(\bar{x},\bar{y})\in \bar{d}} \mathcal{L}_{\text{CE}}\left(f_{\bar{\phi},\bar{\theta}}(\bar{x}), \bar{y}\right) \tag{5}$$

We emphasize that updating the noisy classifier $f_{\bar{\phi},\bar{\theta}}$ every iteration is critical as it can adaptively model the ever-changing noisy label distribution on the fly, where the distribution constantly shifts as the noisy labels are actively corrected to reduce the noise level (See § 3.5).

3.4 Efficient Training

Similar to [98, 44], we propose an efficient training scheme through weight sharing via two-head architecture, as shown in Figure 2. Where the architecture closely resembles the ones of [98, 44], our two-head architecture only shares the feature extractor $\phi = \bar{\phi}$. Unlike the shared feature extractor, our architecture does not share the linear classifier since modeling both noisy and clean data distribution with a single linear classifier is impractical. Based on the two-head architecture, the given samples require only a single inference on the feature extractor for (1) training classifiers, (2) estimating the transition matrix, and (3) correcting labels, which makes model training highly efficient. Thus, we define the clean and noisy classifier as $f_{\phi,\theta}$ and $f_{\phi,\bar{\theta}}$, respectively, to produce our final objective function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{clean}} + \lambda \mathcal{L}_{\text{noisy}} \tag{6}$$

where λ is a loss balancing factor. In order to prevent over-fitting on $\bar{d}$, we introduce λ to the final objective function. We search for the optimal hyperparameter λ for all of our experiments (See Appendix C.7).

Efficiency Analysis. Compared to the vanilla training scheme, which assumes that all labels are clean, we only add a single linear classifier $\bar{\theta}$ with only N additional parameters. Also, our loss only requires a single back-propagation, where the added linear classifier has a negligible computational burden. Our training scheme stands out even more compared to the existing MAML-based methods [105, 121] or multi-stage training [40, 3] (See § 4.2 and Figure 1).

3.5 Label Correction

In this paper, we focus on the efficient, on-the-fly estimation of the label transition matrix to combat label miscorrection. To further demonstrate the effectiveness of our method, we employ a naïve label correction strategy where we feed each noisy set sample $x \in \bar{d}$ to the clean classifier $f_{\phi,\theta}$ to produce a probability vector. If the maximum probability $\max(f_{\phi,\theta}(x))$ is bigger than the threshold ρ, we correct its label to a more probable label. This strategy relies only on the most recent prediction of the model mid-training, so the decision is prone to change. Formally, we can describe the relabeled $\hat{y}$,

$$\hat{y} = \begin{cases} \bar{y}^*, & \text{if } \max(f_{\phi,\theta}(\bar{x})) < \rho \\ \lfloor f_{\phi,\theta}(\bar{x}) / \max(f_{\phi,\theta}(\bar{x})) \rfloor, & \text{otherwise} \end{cases} \tag{7}$$

where $\lfloor \cdot \rfloor$ denotes floor function and $\bar{y}^*$ denotes the original label from $\bar{d}$. $\bar{y}^*$ differs from $\bar{y}$; the former denotes the original label from the noisy dataset, whereas the latter is continuously corrected by the above strategy. Even with this simple strategy, our model shows better performance compared to the state-of-the-art methods. The experimental results suggest that replacing this strategy may further improve the model performance.

4 Experiments

In this section, we evaluate our proposed learning method, FasTEN, in terms of predictive performance (§ 4.1) and efficiency (§ 4.2). We also validate the label correction performance to demonstrate that our method is better in correcting noisy labels (§ 4.3 and Appendix C.6) and experimentally show the robustness of our proposed method towards miscorrected labels. (§ 4.4). We further analyze whether our method successfully estimates the label transition matrix in the case where the label correction shifts the true label transition matrix (§ 4.5) or not (§ 4.6). Additional experimental results and further analyses are described in Appendix C. We provide the source codes[1] for the reproduction of the experiments conducted in this paper.

Baselines using the small clean dataset. We deliberately choose the baselines that utilize the small clean dataset in learning with noisy labels. These baselines are categorized in the following three types. *Re-weighting*: **L2RW** [80] learns to assign weights to training samples based on their gradients. **MW-Net** [85] trains an explicit weighting function with the training samples. **Deep kNN** [3] applies the k-nearest neighbor algorithm to the logit layer of classifiers to find noisy samples. *Label transition matrix estimation*: **GLC** [40] estimates the label transition matrix using the small clean dataset. **MLoC** [101] considers the label transition matrix as trainable parameters to be obtained through meta-learning. *Label correction*: **MLaC** [121] trains a label correction network as a meta-process to provide corrected labels. **MSLC** [105] uses soft labels with loss balancing weight through meta-gradient descent step under the guidance of the clean dataset.

[1] https://github.com/hyperconnect/FasTEN

Table 1: Performance comparison on CIFAR-10/100 datasets under various noise level. Test accuracy (%) with 95% confidence interval of 5-runs is provided.

	Method	Symmetric Noise Level 20%	40%	60%	80%	Asymmetric Noise Level 20%	40%
CIFAR-10	L2RW	88.26 ± 0.79	83.76 ± 0.54	74.54 ± 1.54	42.60 ± 1.71	88.79 ± 0.63	85.86 ± 0.87
	MW-Net	89.76 ± 0.31	86.52 ± 0.28	81.68 ± 0.25	56.56 ± 3.07	91.31 ± 0.25	88.69 ± 0.37
	Deep kNN	90.02 ± 0.35	87.27 ± 0.39	82.80 ± 0.55	68.30 ± 1.21	89.97 ± 0.48	84.56 ± 0.87
	GLC	89.66 ± 0.10	85.30 ± 0.73	80.34 ± 0.73	67.44 ± 1.50	91.56 ± 0.66	**89.76 ± 0.89**
	MLoC	90.50 ± 0.71	87.20 ± 0.35	81.95 ± 0.44	54.64 ± 4.04	91.15 ± 0.16	89.35 ± 0.45
	MLaC	89.75 ± 0.62	86.63 ± 0.56	82.20 ± 0.81	71.94 ± 2.22	91.45 ± 0.32	**90.26 ± 0.48**
	MSLC	90.94 ± 0.45	88.36 ± 0.80	83.93 ± 1.21	64.90 ± 4.84	**91.45 ± 1.35**	89.26 ± 0.52
	FasTEN (ours.)	**91.94 ± 0.28**	**90.07 ± 0.17**	**86.78 ± 0.31**	**79.52 ± 0.78**	**92.29 ± 0.10**	**90.43 ± 0.31**
CIFAR-100	L2RW	57.79 ± 1.88	44.82 ± 4.30	30.01 ± 1.74	10.71 ± 1.79	59.11 ± 2.74	55.12 ± 3.40
	MW-Net	66.73 ± 0.78	59.44 ± 0.91	49.19 ± 1.57	19.04 ± 1.21	67.90 ± 0.78	64.50 ± 0.34
	Deep kNN	59.60 ± 0.97	52.48 ± 1.37	39.90 ± 0.60	23.39 ± 0.75	57.71 ± 0.47	50.23 ± 1.12
	GLC	60.99 ± 0.64	49.00 ± 4.33	33.38 ± 4.09	20.38 ± 1.35	64.43 ± 0.43	54.20 ± 0.86
	MLoC	**68.16 ± 0.41**	62.09 ± 0.33	**54.49 ± 0.92**	20.23 ± 1.86	69.20 ± 0.59	66.48 ± 0.56
	MLaC	49.81 ± 5.59	35.15 ± 5.75	20.15 ± 2.81	12.85 ± 0.87	56.46 ± 3.54	49.20 ± 3.23
	MSLC	**68.62 ± 0.60**	**63.30 ± 0.49**	53.83 ± 0.70	21.07 ± 5.20	**70.86 ± 0.30**	**66.99 ± 0.69**
	FasTEN (ours.)	**68.75 ± 0.60**	**63.82 ± 0.33**	**55.22 ± 0.64**	**37.36 ± 1.15**	**70.35 ± 0.51**	**67.93 ± 0.53**

4.1 Predictive Performance Comparison

CIFAR-10/100 with Synthetic Noise. CIFAR-10/100 [49] have been widely adopted to assess the robustness of the methods to noisy labels. Since CIFAR-10/100 are known as clean datasets, labels are synthetically manipulated to contain noisy labels, injecting two types of noise: symmetric and asymmetric. **Symmetric:** The labels are randomly flipped with uniform distribution. **Asymmetric:** the labels are flipped with class-dependent distribution, following the evaluation protocol of [76, 111]. We claim that most studies report the performance highly overfitted to the test set without hyperparameter tuning on the validation set [105, 53, 73, 74]. Moreover, baseline models employ different backbone networks, making it challenging to dissect the performance improvement whether it originated from each method or the backbone networks. Therefore, we first extract 5K samples as the validation set from the training set containing 50K samples and further extract 1K samples as the clean dataset. Then, we unify the backbone network as ResNet-34 [37], which is widely adopted in various baselines [105, 59]. Note that we do our best to maintain the experimental settings of each method, including the hyperparameters written in the original paper. Detailed settings are deferred to Appendix B.

Results. Table 1 summarizes the evaluation results on CIFAR-10/100. For both CIFAR-10/-100, our proposed FasTEN achieves state-of-the-art performance on various noise levels within 95% confidence intervals. Especially, under a high noise level (80%), our FasTEN considerably outperforms the baselines with small variance on performance, which implies the robustness of our method [52, 51].

Table 2: Test accuracy (%) comparison on Clothing1M dataset with real-world label noise. Rows with † denote results directly borrowed from [121] and ⋆ denotes the result directly borrowed from [56]. All the other results except L2RW [80] are taken from original papers.

	Method	Top-1 Accuracy
Clean set X	Forward⋆	69.91
	T-Revision⋆	70.97
	casualNL	72.24
	IF	72.29
	VolMinNet⋆	72.42
	DivideMix	74.76
	AugDesc	75.11
Clean set O	MLoC	71.10
	L2RW	72.04 ± 0.24
	GLC†	73.69
	MW-Net†	73.72
	MSLC	74.02
	MLaC†	75.78
Ours	FasTEN w/o LC	77.07 ± 0.52
	FasTEN	**77.83 ± 0.17**

These results demonstrate that our proposed method performs well in learning with noisy labels, especially considering its training efficiency (See § 4.2).

Clothing1M with Real-world Noise. Clothing1M [109] is a noisy real-world dataset that consists of one million samples with additional 47K human-annotated clean samples. We use its original splits of clean and noisy data. For a fair comparison, we employ ResNet-50 architecture pretrained with the ImageNet dataset [18] for the initial backbone architecture. Evaluation results on Clothing1M are summarized in Table 2.

Further baselines. We further compare our proposed FasTEN with additional baselines that have already reported their performance on Clothing1M dataset. Since the data split of Clothing1M dataset is the same for all the baselines, we simply obtain the performance of the baselines from their original papers and report the performance in Table 2. **DivideMix** [53] and **AugDesc** [73] leverages semi-supervised learning with various data augmentation strategies. **Forward** [76], **T-Revision** [108], **IF** [45], and **causalNL** [112], and **VolMinNet** [56] are transition matrix estimation methods that use certain data points without clean data points.

Results. As shown in Table 2, our proposed FasTEN achieves remarkable performance on Clothing1M which contains instance-dependent noisy labels, beating the baselines by a large margin. This evaluation result indicates that our proposed FasTEN is more applicable in real-world problems where label corruption frequently occurs, although it does not directly target to address the problem of instance-dependent noisy labels. Similar to previous observations [66, 40], we suspect that using the transition matrix seems to combat instance-dependent noise to some extent. Also, not only that our method shows superior performance over all the baselines that use the small clean set, but it also surpasses the

Table 3: Training time comparison on CIFAR-10 dataset with 80% symmetric noise. Time (hours) per total training on a single RTX 2080Ti GPU are provided with the relative ratio compared to our method.

Method	L2RW [80]	MW-NET [85]	Deep kNN [3]	GLC [40]	MLoC [101]	MLaC [121]	MSLC [105]	**FasTEN (ours.)**
Total Training Time (Relative to Ours.)	4.78 (3.11x)	2.63 (1.71x)	2.32 (1.51x)	2.29 (1.49x)	7.13 (4.64x)	10.2 (6.64x)	2.51 (1.64x)	**1.54**

semi-supervised learning-based methods (DivideMix and AugDesc) without any complex augmentation techniques. Finally, FasTEN shows better performance than T-Revision, causalNL, IF, and VolMinNet, which estimate the transition matrix without the small clean data (this is not a fair comparison). This result indicates that using the small clean data is effective in estimating the transition matrix accurately, leading to performance improvement eventually.

4.2 Training Time Comparison

Setup. To verify the efficiency of our proposed FasTEN, we compare it with the baselines in terms of accuracy by total training time. Total training time is measured on CIFAR-10/-100, respectively, with a single RTX 2080Ti GPU. Test accuracy shows the predictive performance on CIFAR-10/-100 with 20% and 80% symmetric noise ratios, the mildest and most severe noise conditions, respectively. Since Deep kNN and GLC require multiple training stages, the summation of all the hours needed for each training phase is provided.
Results. Figure 1 shows that our FasTEN, which learns the label transition matrix with the single back-propagation in the single-training stage, makes model training more efficient than other baselines that need multiple back-propagation or multiple training stages while showing better performance. Table 3 shows the total training hours of each baseline, including our FasTEN. Our method provides the training time speedup of minimum $\times 1.49$ to maximum $\times 6.64$.

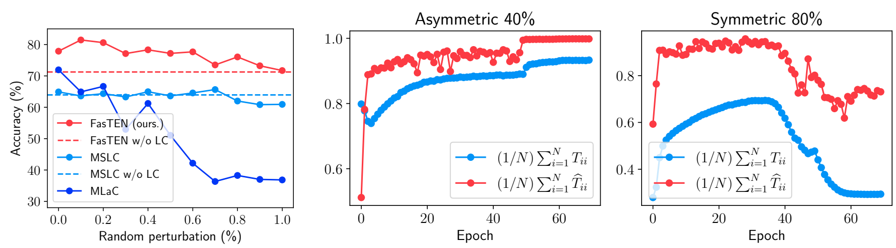
4.3 Label Correction Performance Comparison

We analyze the predictive performance of the baseline methods on all the training samples (Overall) and the wrongly labeled subset of them (Incorrect), respectively. Table 4 demonstrates that our method can successfully correct the noisy labels, where using the label correction further improves the correction performance. This also implies that our FasTEN may be helpful in further cleansing the noisy training set.

We also compare the performance between Overall and Incorrect cases. Reweighting (L2RW, MW-Net, Deep kNN) and transition matrix estimation-based methods (GLC, MLoC) show similar performance between two cases: Overall and Incorrect. However, the performance of the meta-model of MLaC is worse for the Incorrect case, which indicates that the correction from the meta-model

Table 4: Label correction performance comparison on CIFAR-10 with symmetric 80% noise. Accuracy (%) and Negative Log Likelihood (NLL) loss are calculated using the true labels before the synthetic noise is injected. Performance of the trained model on all training samples (Overall) and incorrectly labeled training samples (Incorrect) is measured. † denotes performance extracted from the meta model.

Method		L2RW [80]	MW-Net [85]	Deep kNN [3]	GLC [40]	MLoC [101]	MLaC [121]	MLaC†	MSLC [105]	MSLC†	FasTEN w/o LC	**FasTEN (ours.)**
Acc.	Overall	0.4450	0.6024	0.6471	0.6900	0.6261	0.7567	0.7672	0.6762	0.2821	0.7559	**0.7847**
	Incorrect	0.4447	0.6024	0.6483	0.6903	0.6257	0.7569	0.7382	0.6755	0.2836	0.7560	**0.7861**
NLL	Overall	1.6684	1.6961	1.6085	1.3904	1.7492	0.9868	1.7004	1.2694	1.5989	1.0057	**0.8889**
	Incorrect	1.6674	1.6957	1.6084	1.3881	1.7493	0.9851	1.7299	1.2722	1.5990	1.0033	**0.8877**



(a) Robustness (b) Transition matrix estimation

Fig. 3: (a) Robustness to miscorrected labels on CIFAR-10 with various perturbation strength. Test accuracy (%) of baselines and baselines without the label correction is provided. (b) The plot for the mean of the diagonal term in true transition matrix $\boldsymbol{T}$ and our estimated transition matrix $\widehat{\boldsymbol{T}}$ according to the epoch on CIFAR-10 dataset with symmetric 80% and asymmetric 40% noise.

is less effective where the labels are wrong. Also, notable underperformance of the meta-model of MSLC may indicate the inefficacy of the meta-model. We also analyze the meta-model of the re-weighting methods in the Appendix C.6, where they do not distinguish the wrongly labeled samples well.

4.4 Robustness to Miscorrection: What Happens if Labels are Wrongly Corrected?

This subsection illustrates the robustness of our label correction method to miscorrected labels by comparing it with other label correction methods (MLaC and MSLC) which blindly trust the miscorrected labels as the ground-truth, where we verify the imperfect corrections (See § 4.3). We examine how much this behavior deteriorates the predictive performance.

Setup. We experiment on CIFAR-10 with symmetric 80% noise where there are a maximum number of noisy labels to correct. To simulate the miscorrection, we perturb the corrected labels by injecting artificial noise. We control the degree of random perturbation to observe the robustness of each method on various levels of miscorrection. We further assess the robustness of our FasTEN and MSLC by comparing it with the performance obtained without label correction.
Results. Figure 3a shows our proposed FasTEN outperforms MLaC and MSLC on all the degrees of the random perturbation. MLaC shows steep performance degradation when perturbation worsens, i.e., there are more miscorrected labels. This observation reveals the susceptibility of MLaC. MSLC shows trivial performance gains when labels are corrected, implying that it is not using the full benefits of label correction. Furthermore, when highly perturbed, MSLC performance worsens if it attempts to correct the labels. In contrast, the label correction of our FasTEN improves performance even in harsh situations. FasTEN does not degrade performance even if the correction becomes useless (100% perturbation). These observations show that our FasTEN builds a more robust classifier to miscorrected labels through its efficient estimation of the label transition matrix, acting as a safeguard combating the miscorrected labels.

4.5 On-the-fly Estimation of the Label Transition Matrix

Our proposed FasTEN newly estimates the label transition matrix on every iteration, where the matrix is constantly shifted by label correction. To assess the matrix estimation quality, we compare it with the true label transition matrix.
Setup. We train FasTEN on CIFAR-10 with symmetric 80% and asymmetric 40% noise, which are harsh conditions on symmetric and asymmetric noise injection, respectively. We compare the estimated label transition matrix $\widehat{\boldsymbol{T}}$ with the true label transition matrix $\boldsymbol{T}$ by observing the mean of diagonal term values for each epoch. The mean of the diagonal term in the transition matrix represents the average of the probability that a sample is mapped to a clean label.
Results. Figure 3b shows the overall tendency of the estimated transition matrix (red) to follow the true label matrix (blue). In the asymmetric 40% setting, diagonal term values of the true label transition matrix $\boldsymbol{T}$ gradually increases (blue), which indicates the dataset is cleansed by the label correction. However, in the symmetric 80% case, diagonal term values of the true transition matrix $\boldsymbol{T}$ decreases at the middle of the training. As we maintain the fixed threshold ρ, the total number of corrected samples decreases. Nonetheless, we can conclude that the transition matrix is successfully estimated on shifting noise levels.

Additionally, we observe that the estimated transition matrix $\widehat{\boldsymbol{T}}$ shows higher mean values, i.e., being overconfident on the clean dataset samples. Theoretically, $f_{\phi,\bar{\theta}}$ should correctly approximate the noisy label distribution given enough number of clean samples (See Appendix A.4), but it seems to be overfitting to the clean dataset in practice. This observation is consistent with the popular belief that neural networks tend to learn clean samples first and noisy samples later [1]. For better matrix estimation to yield a more robust classifier [31, 108, 67, 113], it appears that we need to address the overfitting through additional components.

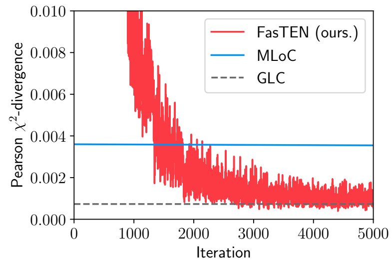


Fig. 4: Plot of transition matrix estimation error for every iteration. Pearson χ^2-divergence of our FasTEN, MLoC and GLC is provided.

4.6 Empirical Convergence Analysis on Estimating the Label Transition Matrix

Setup. This section analyzes the convergence of estimation error between the true label transition matrix $\boldsymbol{T}$ and the estimated transition matrix $\widehat{\boldsymbol{T}}$, comparing our FasTEN to other methods, MLoC and GLC, which learn the transition matrix. For fair comparison, we exclude the label correction for our method.

Results. Figure 4 shows the difference between the probability distribution of the true label transition matrix $\boldsymbol{T}$ and the estimated transition matrix $\widehat{\boldsymbol{T}}$ for each iteration, where Pearson χ^2-divergence is used to measure the discrepancy between the two matrices. GLC error remains fixed (dotted line) as it estimates the transition matrix only once in the entire learning process. The decrease of MLoC error is extremely slow (blue line), implying the high dependence of the initialization of $\widehat{\boldsymbol{T}}$ and its ineffectiveness on estimation. Although our FasTEN does not require multiple stages of training and produces the single mini-batch-based estimate every iteration, it shows fast convergence with a similar estimation error to GLC, which uses all the available data.

5 Conclusion

In this work, we propose a robust and efficient method, FasTEN, which efficiently learns a label transition matrix that mitigates the label miscorrection problem of existing label correction methods. Our proposed FasTEN accurately estimates the label transition matrix using a small clean dataset even if the samples are miscorrected. Moreover, our FasTEN is highly efficient compared to existing methods since it requires single back-propagation through two-head architecture and needs only a single training stage. Extensive experiments show that our method is the fastest and the most robust classifier. Especially, our method achieves remarkable performance on both the real-world noise dataset (Clothing1M) and the synthetic dataset on various noise levels (CIFAR). The detailed analysis shows that our method is robust to miscorrected labels by efficiently estimating the transition matrix shifted by the label correction.

References

1. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: International Conference on Machine Learning. pp. 233–242. PMLR (2017) 13
2. Azadi, S., Feng, J., Jegelka, S., Darrell, T.: Auxiliary image regularization for deep cnns with noisy labels. arXiv preprint arXiv:1511.07069 (2015) 2, 3
3. Bahri, D., Jiang, H., Gupta, M.: Deep k-nn for noisy labels. In: International Conference on Machine Learning. pp. 540–550. PMLR (2020) 2, 4, 7, 8, 11, 12, 25, 30, 35, 36
4. Bartlett, P., Foster, D.J., Telgarsky, M.: Spectrally-normalized margin bounds for neural networks. arXiv preprint arXiv:1706.08498 (2017) 28
5. Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101**(473), 138–156 (2006) 24
6. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: Risk bounds and structural results. Journal of Machine Learning Research **3**(Nov), 463–482 (2002) 28
7. Bekker, A.J., Goldberger, J.: Training deep neural-networks based on unreliable labels. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2682–2686. IEEE (2016) 4, 5, 37
8. Berthon, A., Han, B., Niu, G., Liu, T., Sugiyama, M.: Confidence scores make instance-dependent label-noise learning possible. arXiv preprint arXiv:2001.03772 (2020) 5
9. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities: A nonasymptotic theory of independence. Oxford university press (2013) 28
10. Cao, K., Chen, Y., Lu, J., Arechiga, N., Gaidon, A., Ma, T.: Heteroskedastic and imbalanced deep learning with adaptive regularization. arXiv preprint arXiv:2006.15766 (2020) 3
11. Chang, H.S., Learned-Miller, E., McCallum, A.: Active bias: Training more accurate neural networks by emphasizing high variance samples. arXiv preprint arXiv:1704.07433 (2017) 3, 22
12. Charikar, M., Steinhardt, J., Valiant, G.: Learning from untrusted data. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing. pp. 47–60 (2017) 1, 22, 23, 29
13. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002) 33
14. Chen, P., Liao, B.B., Chen, G., Zhang, S.: Understanding and utilizing deep neural networks trained with noisy labels. In: International Conference on Machine Learning. pp. 1062–1070. PMLR (2019) 1, 3, 22, 23, 24, 32
15. Chen, P., Ye, J., Chen, G., Zhao, J., Heng, P.A.: Robustness of accuracy metric and its inspirations in learning with noisy labels. arXiv preprint arXiv:2012.04193 (2020) 5, 23, 24, 25
16. Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., Liu, Y.: Learning with instance-dependent label noise: A sample sieve approach. In: ICLR (2021) 3
17. Cheng, J., Liu, T., Ramamohanarao, K., Tao, D.: Learning with bounded instance and label-dependent label noise. In: International Conference on Machine Learning. pp. 1789–1799. PMLR (2020) 24

18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 10
19. Drory, A., Avidan, S., Giryes, R.: How do neural networks overcome label noise. arXiv Preprint (2018) 22
20. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: International Conference on Machine Learning. pp. 1126–1135. PMLR (2017) 2, 4, 25, 38
21. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines **30**(4), 681–694 (2020) 1
22. Ghosh, A., Kumar, H., Sastry, P.: Robust loss functions under label noise for deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017) 3
23. Ghosh, A., Lan, A.: Do we really need gold samples for sample weighting under label noise? In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3922–3931 (2021) 2
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014) 1
25. Goldberger, J., Ben-Reuven, E.: Training deep neural-networks using a noise adaptation layer (2016) 3, 4, 5, 25, 37
26. Golowich, N., Rakhlin, A., Shamir, O.: Size-independent sample complexity of neural networks. In: Conference On Learning Theory. pp. 297–299. PMLR (2018) 28
27. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org 1
28. Guan, M., Gulshan, V., Dai, A., Hinton, G.: Who said what: Modeling individual labelers improves classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 38
29. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017) 25, 26
30. Guo, J., Gong, M., Liu, T., Zhang, K., Tao, D.: Ltf: A label transformation framework for correcting label shift. In: International Conference on Machine Learning. pp. 3843–3853. PMLR (2020) 3
31. Han, B., Niu, G., Yu, X., Yao, Q., Xu, M., Tsang, I., Sugiyama, M.: Sigua: Forgetting may make learning with noisy labels more robust. In: International Conference on Machine Learning. pp. 4006–4016. PMLR (2020) 13, 24
32. Han, B., Yao, J., Niu, G., Zhou, M., Tsang, I., Zhang, Y., Sugiyama, M.: Masking: A new perspective of noisy supervision. arXiv preprint arXiv:1805.08193 (2018) 37
33. Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., Sugiyama, M.: Co-teaching: Robust training of deep neural networks with extremely noisy labels. arXiv preprint arXiv:1804.06872 (2018) 22
34. Han, J., Luo, P., Wang, X.: Deep self-learning from noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5138–5147 (2019) 3
35. Han, K., Wang, Y., Xu, Y., Xu, C., Wu, E., Xu, C.: Training binary neural networks through learning with noisy supervision. In: International Conference on Machine Learning. pp. 4017–4026. PMLR (2020) 3

36. Harutyunyan, H., Reing, K., Ver Steeg, G., Galstyan, A.: Improving generalization by controlling label-noise information in neural network weights. In: International Conference on Machine Learning. pp. 4071–4081. PMLR (2020) 3
37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1, 9
38. Hendrycks, D., Lee, K., Mazeika, M.: Using pre-training can improve model robustness and uncertainty. In: International Conference on Machine Learning. pp. 2712–2721. PMLR (2019) 3
39. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. arXiv preprint arXiv:1906.12340 (2019) 3
40. Hendrycks, D., Mazeika, M., Wilson, D., Gimpel, K.: Using trusted data to train deep networks on labels corrupted by severe noise. Advances in neural information processing systems **31** (2018) 1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 22, 25, 26, 34, 36, 38
41. Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., Chang, B.: Disentangling label distribution for long-tailed visual recognition. arXiv preprint arXiv:2012.00321 (2020) 33
42. Hu, W., Li, Z., Yu, D.: Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. arXiv preprint arXiv:1905.11368 (2019) 3
43. Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: beyond empirical risk minimization. Advances in Neural Information Processing Systems **33** (2020) 3
44. Jiang, L., Zhou, Z., Leung, T., Lif, L.J., Fei-Fei, L.: Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In: International Conference on Machine Learning. pp. 2304–2313. PMLR (2018) 3, 4, 7, 22, 25
45. Jiang, Z., Zhou, K., Liu, Z., Li, L., Chen, R., Choi, S.H., Hu, X.: An information fusion approach to learning with instance-dependent label noise. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=ecH2FKaARUp 3, 10
46. Jindal, I., Nokleby, M., Chen, X.: Learning deep networks from noisy labels with dropout regularization. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). pp. 967–972. IEEE (2016) 3
47. Kim, T., Ko, J., Choi, J., Yun, S.Y., et al.: Fine samples for learning with noisy labels. Advances in Neural Information Processing Systems **34** (2021) 3
48. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 30
49. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 3, 9, 30
50. Lee, K.H., He, X., Zhang, L., Yang, L.: Cleannet: Transfer learning for scalable image classifier training with label noise. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5447–5456 (2018) 4, 25
51. Li, D., Chen, C., Liu, W., Lu, T., Gu, N., Chu, S.M.: Mixture-rank matrix approximation for collaborative filtering. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 477–485 (2017) 9
52. Li, D., Chen, C., Lv, Q., Yan, J., Shang, L., Chu, S.: Low-rank matrix approximation with stability. In: International Conference on Machine Learning. pp. 295–303. PMLR (2016) 9
53. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394 (2020) 3, 9, 10

54. Li, J., Wong, Y., Zhao, Q., Kankanhalli, M.S.: Learning to learn from noisy labeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5051–5059 (2019) 4, 25
55. Li, M., Soltanolkotabi, M., Oymak, S.: Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 4313–4324. PMLR (2020) 3
56. Li, X., Liu, T., Han, B., Niu, G., Sugiyama, M.: Provably end-to-end label-noise learning without anchor points. In: International Conference on Machine Learning. PMLR (2021) 4, 10
57. Lienen, J., Hüllermeier, E.: From label smoothing to label relaxation. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI, Online, February 2-9, 2021. AAAI Press (2021) 3
58. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 3, 22
59. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-learning regularization prevents memorization of noisy labels. arXiv preprint arXiv:2007.00151 (2020) 3, 9
60. Liu, T., Tao, D.: Classification with noisy labels by importance reweighting. IEEE Transactions on pattern analysis and machine intelligence **38**(3), 447–461 (2015) 3, 5
61. Liu, Y., Guo, H.: Peer loss functions: Learning from noisy labels without knowing noise rates. In: International Conference on Machine Learning. pp. 6226–6236. PMLR (2020) 3
62. Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S.: Does label smoothing mitigate label noise? In: International Conference on Machine Learning. pp. 6448–6458. PMLR (2020) 3
63. Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., Bailey, J.: Normalized loss functions for deep learning with noisy labels. In: International Conference on Machine Learning. pp. 6543–6553. PMLR (2020) 3
64. Ma, X., Wang, Y., Houle, M.E., Zhou, S., Erfani, S., Xia, S., Wijewickrema, S., Bailey, J.: Dimensionality-driven learning with noisy labels. In: International Conference on Machine Learning. pp. 3355–3364. PMLR (2018) 3, 22
65. Menon, A.K., Rawat, A.S., Reddi, S.J., Kumar, S.: Can gradient clipping mitigate label noise? (2020) 3
66. Menon, A.K., Van Rooyen, B., Natarajan, N.: Learning from binary labels with instance-dependent corruption. arXiv preprint arXiv:1605.00751 (2016) 3, 10, 25
67. Mirzasoleiman, B., Cao, K., Leskovec, J.: Coresets for robust training of deep neural networks against noisy labels. Advances in Neural Information Processing Systems **33** (2020) 2, 3, 4, 13, 22, 23, 24, 37
68. Mnih, V., Hinton, G.E.: Learning to label aerial images from noisy data. In: Proceedings of the 29th International conference on machine learning (ICML-12). pp. 567–574 (2012) 2, 4, 5, 25, 37
69. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of machine learning. MIT press (2018) 24
70. Montgomery-Smith, S.J.: The distribution of rademacher sums. Proceedings of the American Mathematical Society **109**(2), 517–522 (1990) 28
71. Natarajan, N., Dhillon, I.S., Ravikumar, P., Tewari, A.: Learning with noisy labels. In: NIPS. vol. 26, pp. 1196–1204 (2013) 3

72. Neyshabur, B., Bhojanapalli, S., Srebro, N.: A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. arXiv preprint arXiv:1707.09564 (2017) 28
73. Nishi, K., Ding, Y., Rich, A., Höllerer, T.: Augmentation strategies for learning with noisy labels. arXiv preprint arXiv:2103.02130 (2021) 3, 9, 10
74. Ortego, D., Arazo, E., Albert, P., O'Connor, N.E., McGuinness, K.: Multi-objective interpolation training for robustness to label noise. arXiv preprint arXiv:2012.04462 (2020) 3, 9
75. Patrini, G., Nielsen, F., Nock, R., Carioni, M.: Loss factorization, weakly supervised learning and label noise robustness. In: International conference on machine learning. pp. 708–717. PMLR (2016) 2, 3
76. Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L.: Making deep neural networks robust to label noise: A loss correction approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1944–1952 (2017) 3, 4, 5, 7, 9, 10, 25, 37
77. Pleiss, G., Zhang, T., Elenberg, E.R., Weinberger, K.Q.: Identifying mislabeled data using the area under the margin ranking. arXiv preprint arXiv:2001.10528 (2020) 3
78. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 4
79. Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. arXiv preprint arXiv:1412.6596 (2014) 4, 5, 25, 37
80. Ren, M., Zeng, W., Yang, B., Urtasun, R.: Learning to reweight examples for robust deep learning. In: International Conference on Machine Learning. pp. 4334–4343. PMLR (2018) 2, 3, 4, 5, 8, 10, 11, 12, 22, 25, 35, 36
81. Rodrigues, F., Pereira, F.: Deep learning from crowds. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018) 38
82. Scott, C.: A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In: Artificial Intelligence and Statistics. pp. 838–846. PMLR (2015) 5
83. Scott, C., et al.: Calibrated asymmetric surrogate losses. Electronic Journal of Statistics **6**, 958–992 (2012) 24
84. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 761–769 (2016) 3, 22
85. Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D.: Meta-weight-net: Learning an explicit mapping for sample weighting. arXiv preprint arXiv:1902.07379 (2019) 2, 4, 5, 8, 11, 12, 25, 31, 35
86. Song, H., Kim, M., Lee, J.G.: Selfie: Refurbishing unclean samples for robust deep learning. In: International Conference on Machine Learning. pp. 5907–5915. PMLR (2019) 3, 22
87. Song, H., Kim, M., Park, D., Lee, J.G.: How does early stopping help generalization against label noise? arXiv preprint arXiv:1911.08059 (2019) 3
88. Sukhbaatar, S., Bruna, J., Paluri, M., Bourdev, L., Fergus, R.: Training convolutional networks with noisy labels. arXiv preprint arXiv:1406.2080 (2014) 4, 5, 7, 25, 37

89. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014) 1
90. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5552–5560 (2018) 3
91. Tanno, R., Saeedi, A., Sankaranarayanan, S., Alexander, D.C., Silberman, N.: Learning from noisy labels by regularized estimation of annotator confusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11244–11253 (2019) 38
92. Thekumparampil, K.K., Khetan, A., Lin, Z., Oh, S.: Robustness of conditional gans to noisy labels. arXiv preprint arXiv:1811.03205 (2018) 25
93. Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., Mohd-Yusof, J.: Combating label noise in deep learning using abstention. arXiv preprint arXiv:1905.10964 (2019) 3
94. Van Rooyen, B., Menon, A.K., Williamson, R.C.: Learning with symmetric label noise: The importance of being unhinged. arXiv preprint arXiv:1505.07634 (2015) 2, 3
95. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013) 23, 24
96. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999) 23
97. Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S.: Learning from noisy large-scale datasets with minimal supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 839–847 (2017) 4, 25
98. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. arXiv preprint arXiv:1606.04080 (2016) 7
99. Wang, Y., Ma, X., Chen, Z., Luo, Y., Yi, J., Bailey, J.: Symmetric cross entropy for robust learning with noisy labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 322–330 (2019) 3
100. Wang, Y., Kucukelbir, A., Blei, D.M.: Robust probabilistic modeling with bayesian data reweighting. In: International Conference on Machine Learning. pp. 3646–3655. PMLR (2017) 3
101. Wang, Z., Hu, G., Hu, Q.: Training noise-robust deep neural networks via meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4524–4533 (2020) 2, 4, 8, 11, 12, 25, 34, 35, 38
102. Wang, Z., Zhu, H., Dong, Z., He, X., Huang, S.L.: Less is better: Unweighted data subsampling via influence function. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6340–6347 (2020) 3
103. Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., Liu, Y.: Learning with noisy labels revisited: A study using real-world human annotations. In: ICLR (2022) 31
104. Wu, P., Zheng, S., Goswami, M., Metaxas, D., Chen, C.: A topological filter for learning with label noise. arXiv preprint arXiv:2012.04835 (2020) 3
105. Wu, Y., Shu, J., Xie, Q., Zhao, Q., Meng, D.: Learning to purify noisy labels via meta soft label corrector. arXiv preprint arXiv:2008.00627 (2020) 2, 3, 4, 5, 7, 8, 9, 11, 12, 22, 25, 31, 33
106. Xia, X., Liu, T., Han, B., Gong, M., Yu, J., Niu, G., Sugiyama, M.: Sample selection with uncertainty of losses for learning with noisy labels. arXiv preprint arXiv:2106.00445 (2021) 3

107. Xia, X., Liu, T., Han, B., Wang, N., Gong, M., Liu, H., Niu, G., Tao, D., Sugiyama, M.: Part-dependent label noise: Towards instance-dependent label noise. Advances in Neural Information Processing Systems **33** (2020) 3, 5
108. Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., Sugiyama, M.: Are anchor points really indispensable in label-noise learning? arXiv preprint arXiv:1906.00189 (2019) 2, 4, 5, 6, 7, 10, 13, 23, 24, 25, 28, 37
109. Xiao, T., Xia, T., Yang, Y., Huang, C., Wang, X.: Learning from massive noisy labeled data for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2691–2699 (2015) 3, 10, 31
110. Yang, F., Koyejo, S.: On the consistency of top-k surrogate losses. In: International Conference on Machine Learning. pp. 10727–10735. PMLR (2020) 24
111. Yao, J., Wu, H., Zhang, Y., Tsang, I.W., Sun, J.: Safeguarded dynamic label regression for noisy supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9103–9110 (2019) 4, 9
112. Yao, Y., Liu, T., Gong, M., Han, B., Niu, G., Zhang, K.: Instance-dependent label-noise learning under a structural causal model. Advances in Neural Information Processing Systems **34** (2021) 3, 10
113. Yao, Y., Liu, T., Han, B., Gong, M., Deng, J., Niu, G., Sugiyama, M.: Dual t: Reducing estimation error for transition matrix in label-noise learning. arXiv preprint arXiv:2006.07805 (2020) 2, 4, 5, 7, 13, 23, 24, 37
114. Yi, K., Wu, J.: Probabilistic end-to-end noise correction for learning with noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7017–7025 (2019) 3
115. Yu, X., Liu, T., Gong, M., Tao, D.: Learning with biased complementary labels. In: Proceedings of the European conference on computer vision (ECCV). pp. 68–83 (2018) 25
116. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 (2016) 23
117. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 3
118. Zhang, X., Wu, X., Chen, F., Zhao, L., Lu, C.T.: Self-paced robust learning for leveraging clean labels in noisy data. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6853–6860 (2020) 4, 25
119. Zhang, Y., Niu, G., Sugiyama, M.: Learning noise transition matrix from only noisy labels via total variation regularization. In: International Conference on Machine Learning. pp. 12501–12512. PMLR (2021) 3
120. Zhang, Z., Sabuncu, M.R.: Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836 (2018) 3
121. Zheng, G., Awadallah, A.H., Dumais, S.: Meta label correction for noisy label learning. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence (2021) 1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 22, 25, 33
122. Zheng, S., Wu, P., Goswami, A., Goswami, M., Metaxas, D., Chen, C.: Error-bounded correction of noisy labels. In: International Conference on Machine Learning. pp. 11447–11457. PMLR (2020) 3
123. Zhu, Z., Liu, T., Liu, Y.: A second-order approach to learning with instance-dependent label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10113–10123 (2021) 3