

# TRAINREF: CURATING DATA WITH LABEL DISTRIBUTION AND MINIMAL REFERENCE FOR ACCURATE PREDICTION AND RELIABLE CONFIDENCE

Murong Ma<sup>1\*</sup>, Ruofan Liu<sup>1</sup>, Yun Lin<sup>2†</sup>, Zhiyong Huang<sup>1</sup>, Jin Song Dong<sup>1</sup>

<sup>1</sup>School of Computing, National University of Singapore

<sup>2</sup>School of Computer Science, Shanghai Jiao Tong University

## ABSTRACT

Practical classification requires both high predictive accuracy and reliable confidence for human-AI collaboration. Given that a high-quality dataset is expensive and sometimes impossible, learning with noisy labels (LNL) is of great importance. The state-of-the-art works propose many denoising approaches by categorically correcting the label noise, i.e., change a label from one class to another, which can suffer from *normality pollution* and *class ambiguity*. The normality pollution happens when the noise ratio gets higher, leading to prediction inaccuracy, as such approaches intrinsically learn normality from the noisy dataset. The class ambiguity happens when the number of classes increases, leading to less reliable prediction confidence.

In this work, we propose a training-time data-curation framework, TrainRef, to uniformly address prediction accuracy and confidence calibration by (1) an extrinsic small set of reference samples  $\mathcal{D}_{\text{ref}}$  to avoid normality pollution and (2) curate labels into a class distribution instead of a categorical class to handle sample ambiguity. Our insights lie in that the extrinsic information allows us to select more precise clean samples even when  $|\mathcal{D}_{\text{ref}}|$  equals to the number of classes (i.e., one sample per class). Technically, we design (1) a reference augmentation technique to select clean samples from the dataset based on  $\mathcal{D}_{\text{ref}}$ ; and (2) a model-dataset co-evolving technique for a near-perfect embedding space, which is used to vote on the class-distribution for the label of a noisy sample. Extensive experiments on CIFAR-100, Animal10N, and WebVision demonstrate that TrainRef outperform the state-of-the-art denoising techniques (DISC, L2B, and DivideMix) and model calibration techniques (label smoothing, Mixup, and temperature scaling). Furthermore, our user study shows that the model confidence trained by TrainRef well aligns with human intuition. More demonstration, proof, and experimental details are available at <https://sites.google.com/view/train-ref>.

## 1 INTRODUCTION

Practical classification application, such as medical diagnosis (Rosenbacke et al., 2024), autonomous driving (Delavari et al., 2025), and fraud detection (Perini & Davis, 2023), requires both accuracy and reliable confidence. Recent work by Kalai et al. (Kalai et al., 2025) shows that even highly capable LLMs tend to produce overconfident false predictions (hallucinations), emphasizing the importance of calibration as a peer to accuracy. The confidence is useful for the model users to decide when to adopt model decision (Corbière et al., 2019; Pan et al., 2020). Such model performance (both accuracy and confidence) usually requires high-quality datasets, however, which are usually expensive, sometimes impossible (AI, 2024; Forbes, 2024).

Therefore, learning with noisy labels (LNL) solutions emerge to address the challenge. The solutions evolve from label transition matrix (Hendrycks et al., 2018; Patrini et al., 2017), sample-reweighting techniques (Li et al., 2020; Sheng et al., 2024; Liu et al., 2020), gravitating to the pseudo-labeling

\*Part of this work was conducted during a visit to SJTU.

†Corresponding author: [lin\\_yun@sjtu.edu.cn](mailto:lin_yun@sjtu.edu.cn)

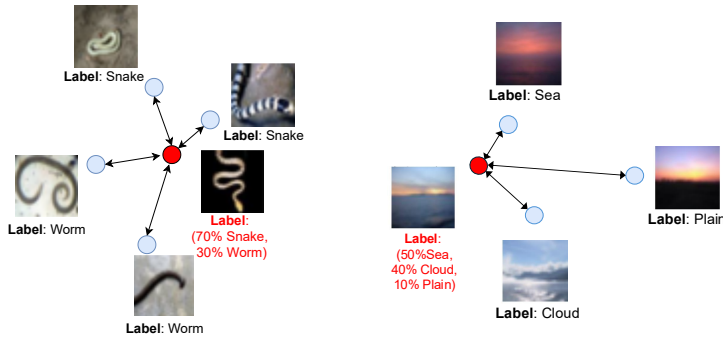


Figure 1: Examples of ambiguous samples from CIFAR-100 where the ground-truth label is distributional rather than categorical. (Left) Snake vs. Worm, where the label can be 70% Snake and 30% Worm. (Right) Sea vs. Cloud vs. Plain, where the label may be distributed as 50% Sea, 40% Cloud, and 10% Plain. Categorical labels for such cases encourage overconfident predictions, whereas distributional labels better capture inherent ambiguity.

techniques which curate the noisy data from one class to another via semi-supervised learning (Kim et al., 2021; Nishi et al., 2021; Lu et al., 2021; Zheltonozhskii et al., 2022; Li et al., 2023). While generally effective in improving predictive accuracy, the performance of such categorical curation (i.e., change one label with another) to learn reliable confidence is limited, especially when the number of classes grows, leading to more ambiguous samples (see Figure 1). In practice, to classify a clinical note into an ICD code, there can be more than 10K classes. In addition, those solutions curate the labels by learning the label normalities from the *intrinsic* noisy dataset. When the noise ratio is high enough, the noisy samples can form the *polluting* normalities, leading to mis-curation. On the other hand, confidence calibration solutions include label smoothing (Müller et al., 2019), mixup (Zhang, 2017), and temperature scaling (Hinton, 2015; Guo et al., 2017) which mitigates prediction overconfidence for more reliable class distributions. However, their performance is limited in addressing the mis-confidence caused by label noise.

In this work, we propose a training-time data-curation solution, TrainRef, with the following features.

- **Distributional Curation:** Different from categorical curation, we learn distributional curation during training, i.e., change a label from one class to a class distribution, for having high predictive accuracy and reliable confidence in a uniform manner.
- **Extrinsic Reference Set:** Different from the state-of-the-arts (Li et al., 2020; Sheng et al., 2024; Liu et al., 2020; Li et al., 2023) which curate the labels by learning the label normalities from the *intrinsic* noisy dataset, our approach introduces a tiny trusted set of reference samples with the ground-truth label,  $\mathcal{D}_{ref}$ , as label normalities, to avoid normality pollution caused by high noise ratios. Our approach can be effective even when  $|\mathcal{D}_{ref}|$  is as small as the number of classes (one sample per class).

Given a trusted reference set  $\mathcal{D}_{ref}$  and a noisy dataset  $\tilde{\mathcal{D}}$ , TrainRef adopts a three-stage training routine by co-evolving the model embedding space and the curated dataset. First, we obtain a label-agnostic encoder  $\theta$  by pre-training masked-image modelling (MIM) (Pathak et al., 2016; Peng et al., 2022) task on  $\tilde{\mathcal{D}}$ . As a result, the embedding space of  $\theta$  is immune to noise. Next, we design a reference augmentation technique to select clean samples  $\hat{\mathcal{D}}(\hat{\mathcal{D}} \subset \tilde{\mathcal{D}})$  based on  $\mathcal{D}_{ref}$ , regarding the agreement between the samples in  $\mathcal{D}_{ref}$  and the samples in  $\tilde{\mathcal{D}}$  through the influence functions (Koh & Liang, 2017; Pruthi et al., 2020; ?) computed on  $\theta$ . Finally, we co-evolve the model  $\theta^*$  and the dataset  $\hat{\mathcal{D}}$  by iteratively (1) fine-tuning a model  $\theta^*$  with learned clean dataset  $\hat{\mathcal{D}}$  and (2) curating and enhancing the dataset  $\hat{\mathcal{D}}$  by voting the label distribution for noisy samples with their neighbouring clean samples on the embedding space of  $\theta^*$ .

Our extensive experiments show that TrainRef significantly outperform the state-of-the-art denoising solutions (i.e., L2B (Zhou et al., 2024), DISC (Li et al., 2023), LSL (Kim et al., 2024)) by 1.82% to 8.19% across various benchmarks, and the state-of-the-art confidence calibration solutions (i.e., label smoothing (Müller et al., 2019), mixup (Zhang, 2017), and temperature scaling (Hinton, 2015;

Guo et al., 2017)) by consistently achieving lower ECE indicating better calibration. Furthermore, in a blind user study, participants agree with TrainRef’s confidence estimates over previous SOTA (DISC + mixup) in **75%** of cases, corroborating its reliability in practice. More demonstration, proof, and experimental details are available at <https://sites.google.com/view/train-ref>.

## 2 PROBLEM STATEMENT

Given a collected dataset  $\tilde{\mathcal{D}} = \{(\mathbf{x}, \tilde{\mathbf{y}})\}_{i=1}^N$  where each sample consists of an input  $\mathbf{x} \in \mathbb{R}^d$  and a one-hot categorical label  $\tilde{\mathbf{y}} = [y_1, y_2, \dots, y_C]$ ,  $y_c \in \{0, 1\}$ ,  $C$  is the number of classes. We assume that the true label is a class distribution instead of a one-hot class, that is,  $\mathcal{D}^* = \{(\mathbf{x}, \mathbf{y}^*)\}$ , where  $\mathbf{y}^* = [y_1^*, y_2^*, \dots, y_C^*]$ ,  $y_c \in [0, 1]$ ,  $\sum_{i=1}^C y_c^* = 1$ . Thus, we define two types of label misinformation:

- **Categorical Noise:** The ground-truth label  $\mathbf{y}^*$  is one-hot (e.g., a 0-entropy distributional label) but  $\arg \max_c y_c^* \neq \arg \max_c \tilde{y}_c$ .
- **Distributional Noise:** The ground-truth label  $\mathbf{y}^*$  is a Dirichlet distribution over  $C$  classes, and  $\mathbf{y}^* \neq \tilde{\mathbf{y}}$ . Intuitively, such noise can lead to an over- or under-confident model.

Our goal is to learn a parameterized mapping  $f_\theta(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^h$  that projects each input into an embedding space for downstream tasks. We obtain our estimator  $\hat{\theta}$  by minimizing the empirical risk over a finite dataset with distributional labels (Equation 1). Here,  $\lambda \|f_\theta\|$  denotes a regularization term, such as weight decay, that penalizes large parameter values to control the model’s capacity.

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{N} \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \tilde{\mathcal{D}}} [\mathcal{L}(f_\theta(\mathbf{x}_i), \tilde{\mathbf{y}}_i) + \lambda \|f_\theta\|] \quad (1)$$

In the limit as  $N \rightarrow \infty$ , a well-behaved estimator  $\hat{\theta}$  should converge to the true risk minimizer.

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(\mathbf{x}_i, \mathbf{y}_i^*) \sim \mathcal{D}^*} [\mathcal{L}(f_\theta(\mathbf{x}_i), \mathbf{y}_i^*)] + \lambda \|f_\theta\| \quad (2)$$

However, in practice, this asymptotic consistency can break down when the dataset  $\tilde{\mathcal{D}}$  contains label misinformation (i.e., both categorical and distributional noise).

**Rationale** If the optimal embedding  $f_{\theta^*}(\cdot)$  were available, and  $\hat{\mathbf{y}} = g \circ f$ , where  $g(\cdot) : \mathbb{R}^h \rightarrow \mathbb{R}^C$  is the classification head. Then by Representer Point Theorem (Schölkopf et al., 2001), any model prediction at query sample  $\mathbf{x}_t$  can be expressed as a linear combination of representative samples  $\mathbf{x}_i$ , weighted by their similarity  $k(\mathbf{x}_t, \mathbf{x}_i)$  (Equation 3), where each coefficient  $\alpha_i$  depends only on the representative sample  $(\mathbf{x}_i, \tilde{\mathbf{y}}_i)$ .

$$\hat{\mathbf{y}}(\mathbf{x}_t) = \sum_{(\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \mathcal{D}_{ref}} \alpha_i \cdot k(\mathbf{x}_i, \mathbf{x}_t) \quad (3)$$

In TrainRef, (i) we estimate the optimal embedding function  $\hat{f}$ , and (ii) we collect a clean reference set with diverse class prototypes,  $\mathcal{D}_{ref} = \{(\mathbf{x}_{ref}, \mathbf{y}_{ref})\}$ . Then given a sample  $(\mathbf{x}, \tilde{\mathbf{y}})$ , we can curate the label as Equation 4.

$$\hat{\mathbf{y}}^*(\mathbf{x}_t) = \frac{1}{|\mathcal{D}_{ref}|} \sum_{(\mathbf{x}_{ref}, \mathbf{y}_{ref}) \in \mathcal{D}_{ref}} \mathbf{y}_{ref} \cdot k(\hat{f}(\mathbf{x}_{ref}), \hat{f}(\mathbf{x}_t)) \quad (4)$$

Here, we set  $\alpha_i = \frac{\mathbf{y}_{ref}}{|\mathcal{D}_{ref}|}$ , i.e. each clean label directly votes in proportion to its similarity with the query. We construct the curated dataset  $\hat{\mathcal{D}} = \{(\mathbf{x}_t, \hat{\mathbf{y}}^*(\mathbf{x}_t)) | \mathbf{x}_t \in \tilde{\mathcal{D}}\}$  to replace  $\tilde{\mathcal{D}}$  in the follow-up training processes, and we hope the  $\hat{\theta}$  learnt from  $\hat{\mathcal{D}}$  is closer to  $\theta^*$ .

**Practical Challenges.** Applying Equation 4 in practice requires overcoming two challenges:

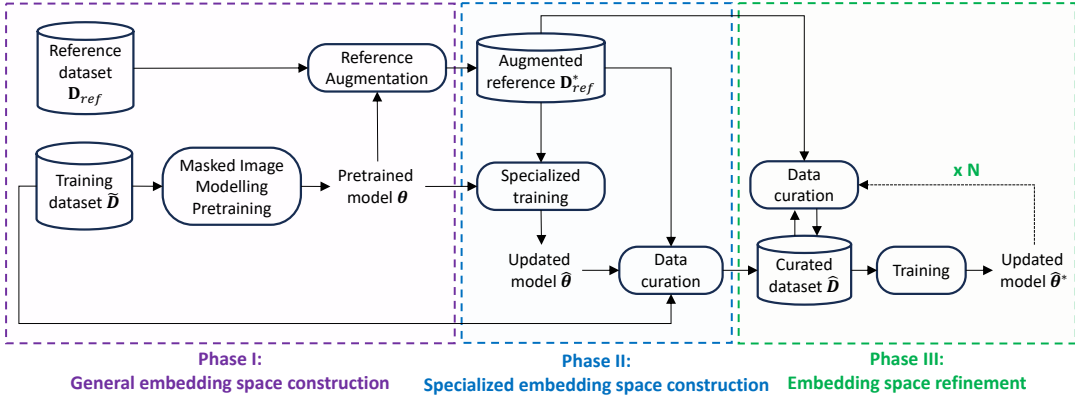


Figure 2: Overview of TrainRef: given a noisy training dataset  $\tilde{\mathcal{D}}$  and a small reference dataset  $\mathcal{D}_{ref}$ , we (1) explore a near-perfect embedding space and (2) augment the reference to learn a function parameterized by  $\hat{\theta}^*$  to minimize the empirical risk. The learned embedding space and the curated dataset are mutually influenced to converge.

- **Accurate embedding-space estimation.** Even with heuristics like the memorization effect (Liu et al., 2020) or confidence-based selection (Li et al., 2020; 2023), noisy or over-confident labels in  $\tilde{\mathcal{D}}$  can skew the learned embedding space when minimizing the empirical risk (Eq. 1).
- **Reference-set diversity.** Manually verified clean references are typically few. We therefore need methods for augmenting and measuring diversity so that our reference pool contains sufficiently representative prototypes for reliable label voting.

Next, we introduce our solution TrainRef to address the above challenges.

### 3 APPROACH

Figure 2 illustrates the overall framework of TrainRef, which learns a deep classifier from a dataset  $\tilde{\mathcal{D}}$  with both categorical and distributional noise, aided by a small clean reference set  $\mathcal{D}_{ref}$ . The core idea is to construct a reliable embedding space  $\mathcal{S}^*$  and an augmented reference set  $\mathcal{D}_{ref}^*$ , such that samples in  $\mathcal{D}_{ref}^*$  can collectively provide accurate supervision for the rest of the dataset via neighborhood voting. The training procedure consists of three iterative phases:

**Phase I: General Embedding Space Construction.** In this stage, we pre-train model  $f_\theta$  from the training dataset  $\tilde{\mathcal{D}}$  with loss functions independent from the labels. Specifically, we choose MIM (Masked Image Modelling) (Pathak et al., 2016; Peng et al., 2022) as the pretraining task which aims to recover a patch in an image sample from its neighboring patches. By this means, we learn an initial embedding space, allowing us to extract *general* semantically meaningful features from the samples.

**Phase II: Special Embedding Space Construction.** In the next stage, we augment the small reference set  $\mathcal{D}_{ref}$  by introducing a subset of *clean* samples in  $\tilde{\mathcal{D}}$ . We consider a sample  $(\mathbf{x}, \mathbf{y})$  in  $\tilde{\mathcal{D}}$  as clean if  $(\mathbf{x}, \mathbf{y})$  is consistent with all the samples in  $\mathcal{D}_{ref}$ . In this work, we introduce influence function (Koh & Liang, 2017) to measure such consistency (see Section 3.1). As a result, we have an augmented reference set  $\mathcal{D}_{ref}^*$  which can be used to further retrain  $\theta$  to  $\hat{\theta}$ . Comparing to  $\theta$ , the embedding space of  $\hat{\theta}$  is more classification-relevant.

**Phase III: Embedding Space Refinement.** In the final stage, we iteratively update the learned model  $\hat{\theta}^*$  and its embedding space and the curated model  $\hat{\mathcal{D}}$ . On one hand, we curate the dataset  $\tilde{\mathcal{D}}$  to  $\hat{\mathcal{D}}$  based on the learned embedding space and a set of discriminated reference samples. On the other hand, the curated dataset  $\hat{\mathcal{D}}$  is further used to update the model to  $\hat{\theta}^*$ . The co-evolution process terminates once both the refined labels  $\hat{\mathbf{y}}^*$  and model parameters  $\hat{\theta}$  converge (see Section 3.2).

Given the space limit, readers can refer to the Appendix A for the dictionary of symbols.

### 3.1 REFERENCE AUGMENTATION

Our insight lies in that noisy samples are likely to generate strong conflicting training signals with the ground-truth references, and clean samples can generate aligning or negligible training signals.

Technically, given a pre-trained model  $f_\theta$  from masked image modeling, we append a fully connected classification head  $g_\phi(\cdot) : \mathbb{R}^h \rightarrow \mathbb{R}^C$  parameterized by  $\phi$ , which is fine-tuned on  $\mathcal{D}_{\text{ref}}$ . The layer is used to measure the agreement between a target training sample  $s_i = (\mathbf{x}_i, \tilde{\mathbf{y}}_i) \in \tilde{\mathcal{D}}$  and any reference samples  $\mathbf{s}_{\text{ref}} = (\mathbf{x}_{\text{ref}}, \mathbf{y}_{\text{ref}}) \in \mathcal{D}_{\text{ref}}$ . In this work, we adopt TraceIn (Pruthi et al., 2020), a practical influence function measuring how likely fitting a training sample  $s_i$  is helpful or harmful to predict a reference sample  $\mathbf{s}_{\text{ref}}$ .

$$IF(\mathbf{s}_i, \mathbf{s}_{\text{ref}}) = \frac{\nabla_\phi \mathcal{L}_i^\top \nabla_\phi \mathcal{L}_{\text{ref}}}{\|\nabla_\phi \mathcal{L}_i\| \cdot \|\nabla_\phi \mathcal{L}_{\text{ref}}\|}, \quad \text{where} \quad \begin{cases} \mathcal{L}_i = \mathcal{L}(g_\phi \circ f_\theta(\mathbf{x}_i), \tilde{\mathbf{y}}_i) \\ \mathcal{L}_{\text{ref}} = \mathcal{L}(g_\phi \circ f_\theta(\mathbf{x}_{\text{ref}}), \mathbf{y}_{\text{ref}}) \end{cases} \quad (5)$$

$$IF(\mathbf{s}_i, \mathcal{D}_{\text{ref}}) = \frac{1}{T \cdot |\mathcal{M}_i|} \sum_{t=1}^T \sum_{\mathbf{s}_{\text{ref}} \in \mathcal{M}_i} IF(\mathbf{s}_i, \mathbf{s}_{\text{ref}}), \quad \text{where} \quad \mathcal{M}_i = \{(\mathbf{x}_{\text{ref}}, \mathbf{y}_{\text{ref}}) \in \mathcal{D}_{\text{ref}} \mid \mathbf{y}_{\text{ref}} = \tilde{\mathbf{y}}_i\} \quad (6)$$

As shown in Equation 6, the influence of training sample  $i$  is calculated as the gradient alignment between this sample and all reference samples with the same label as  $\tilde{\mathbf{y}}_i$ . The final influence score is averaged over  $T$  training checkpoints. Training samples with positive  $IF(\mathbf{s}_i, \mathcal{D}_{\text{ref}})$  larger than threshold  $\delta_{IF}$  are used to construct a larger augmented reference set  $\mathcal{D}_{\text{ref}}^*$ .  $\delta_{IF}$  is tuned to be 0.8, and we demonstrate its robustness with different settings in the Appendix F.3. The augmented reference set is further used to update  $\theta$  and  $\phi$ , collectively referred to as  $\hat{\theta}$ .

### 3.2 CURATION-TRAINING CO-EVOLUTION

We formalize co-evolution as an alternating optimization scheme closely analogous to the Expectation–Maximization (EM) algorithm (Dempster et al., 1977), composed of the following two phases:

- **Curation (C-step):** Holding the current embedding estimator  $\hat{f}$  fixed, we apply Equation 4 to compute refined curated labels  $\hat{\mathbf{y}}^*$ . In this phase, each sample is assigned a “responsibility” weight over  $C$  classes, analogous to the E-step in EM algorithm (Dempster et al., 1977), i.e., posterior mixture-component assignments.
- **Parameter update (P-step):** Using the curated labels  $\hat{\mathbf{y}}^*$  as targets, we update  $f_\theta$  by empirical risk minimization. This update refines the embedding space, which in turn yields more accurate similarity estimation  $k(\hat{f}(\mathbf{x}_{\text{ref}}), \hat{f}(\mathbf{x}_i))$  for the subsequent C-step, analogous to the M-step in EM algorithm (Dempster et al., 1977), i.e., parameter re-estimation given fixed responsibilities.

By iterating these two phases, the embedding function and the curated labels co-adapt dynamically: improved embeddings produce more reliable pseudo-labels, and more accurate pseudo-labels guide sharper embeddings, thereby ensuring progressive convergence despite the presence of label noise.

**C-step Curation Principle.** We estimate refined labels  $\hat{\mathbf{y}}^*$  for all samples in the noisy dataset  $\tilde{\mathcal{D}}$ , based on neighborhood voting from the augmented reference set  $\mathcal{D}_{\text{ref}}^*$ . Given a sample  $(\mathbf{x}, \tilde{\mathbf{y}})$ , the estimated label distribution  $\hat{\mathbf{y}}$  is defined as:

$$\hat{\mathbf{y}}(\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \sum_{(\mathbf{x}_{\text{ref}}, \mathbf{y}_{\text{ref}}) \in \mathcal{D}_{\text{ref}}^*} \mathbf{y}_{\text{ref}} \cdot \mathbb{1}(\mathbf{x}_{\text{ref}} \in \mathcal{D}_{\text{vote}}^*(\mathbf{x})) \cdot \text{Cosine}(f(\mathbf{x}), f(\mathbf{x}_{\text{ref}})) \quad (7)$$

where we use cosine similarity as the kernel function, and  $Z(\mathbf{x})$  is a normalization constant ensuring  $\hat{\mathbf{y}}$  forms a valid probability distribution:  $Z(\mathbf{x}) = \sum_{\mathbb{1}(\mathbf{x}_{\text{ref}} \in \mathcal{D}_{\text{vote}}^*(\mathbf{x}))} \text{Cosine}(f(\mathbf{x}), f(\mathbf{x}_{\text{ref}}))$ .

To mitigate voting noise, we restrict the voting pool using an indicator function  $\mathbb{1}(\mathbf{x}_{\text{ref}} \in \mathcal{D}_{\text{vote}}^*(\mathbf{x}))$  that enforces two criteria: (i) *semantic relevance* to the query sample, defined as

$$\mathcal{D}_{\text{vote}}(\mathbf{x}) = \{(\mathbf{x}_{\text{ref}}, \mathbf{y}_{\text{ref}}) \in \mathcal{D}_{\text{ref}}^* \mid \text{Cosine}(f(\mathbf{x}), f(\mathbf{x}_{\text{ref}})) \geq \tau\} \quad (8)$$

, where  $\tau$  is set to the 75th percentile of the cosine-similarity distribution, and (ii) *inter-diversity* among selected references. For the latter, we construct a subset of  $k$  samples ( $k$  is set to half the pool size) that covers the distribution of  $\mathcal{D}_{\text{vote}}$ :

$$\mathcal{D}_{\text{vote}}^*(\mathbf{x}) = \arg \max_{S \subseteq \mathcal{D}_{\text{vote}}(\mathbf{x}), |S|=k} \sum_{i < j} \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|^2 \quad (9)$$

To sharpen the refined label distribution and emphasize confident predictions, we apply a temperature-controlled transformation following DivideMix (Li et al., 2020). Given  $t \in (0, 1]$ , the sharpened label is computed as  $\hat{\mathbf{y}}^* = \hat{\mathbf{y}}^{1/t} / (\mathbf{1}^\top \hat{\mathbf{y}}^{1/t})$ , with  $t = 0.5$  in all experiments.

**P-step Optimization Objective.** Given the current model parameters  $\theta$  and a curated dataset  $\hat{\mathcal{D}} = (\mathbf{x}_i, \hat{\mathbf{y}}_i^*)$  obtained from the C-step, we update  $\theta$  by minimizing the following objective:

$$\hat{\theta} = \arg \min_{\theta, \phi} \frac{1}{|\hat{\mathcal{D}}|} \sum_{i=1}^{|\hat{\mathcal{D}}|} [\mathcal{L}_{\text{CE}}(g_\phi \circ f_\theta(\mathbf{x}_i), \hat{\mathbf{y}}_i^*) + \lambda \|\theta\|_2^2] \quad (10)$$

Here,  $\mathcal{L}_{\text{CE}}(\cdot, \cdot)$  denotes the cross-entropy loss between the model’s prediction and the curated label, and  $\|\theta\|_2^2$  is an  $\ell_2$  regularization term that discourages overfitting and promotes smoother parameter. We provide theoretical proof of convergence on our website (TrainRef, 2025).

## 4 EXPERIMENT

We evaluate our approach with the following research questions, each addressed with an experiment:

- **RQ1 (Predictive Accuracy):** How effectively does TrainRef achieve predictive accuracy compared to state-of-the-art label denoising methods?
- **RQ2 (Confidence Reliability):** How effectively does TrainRef improve confidence calibration compared to state-of-the-art calibration methods?
- **RQ3 (User study):** In practical use, to what extent do humans agree with the confidence estimates produced by TrainRef?
- **RQ4 (Ablation study):** What are the contributions of each component to the overall performance?

Implementation details and qualitative analyses are deferred to Appendix C due to space limits.

### 4.1 RQ1: PREDICTIVE ACCURACY

**Synthetic Noisy Datasets.** We evaluate TrainRef on CIFAR-100 (Krizhevsky et al., 2009) under three common noise types: (1) *instance-dependent noise* (IDN) (Xia et al., 2020), where each instance is assigned a noise rate from a truncated Gaussian distribution with class-level rates chosen randomly; (2) *symmetric noise* (Sym.) (Li et al., 2020), where labels are flipped uniformly at random to any other class; and (3) *asymmetric noise* (Asym.), where labels are flipped to semantically similar or neighboring classes at a fixed rate. Following Li et al. (2023), we set symmetric noise levels to  $\rho \in \{20\%, 50\%, 80\%\}$  and both asymmetric and instance-dependent noise to  $\rho = 40\%$ .

**Real-World Noisy Datasets.** To assess TrainRef in practical settings, we evaluate it on Web-Vision1.0 (Li et al., 2017) and Animal-10N (Song et al., 2019). WebVision1.0 contains 2.4M web-crawled images from Google and Flickr. Animal-10N consists of noisy labels from five pairs of visually similar animal species. Both are challenging real-world benchmarks.

**Results on CIFAR with synthetic noise.** Table 1 shows the generalization performance under various noise levels on the CIFAR-100 dataset. Overall, TrainRef consistently outperforms all the baselines, serving as a new state-of-the-art. Specifically, under severe noise setting (e.g., 80% symmetric), TrainRef surpasses L2B-C2D by over 10% and remains robust in challenging cases like asymmetric (40%) and instance-dependent (40%) noise, outperforming DISC (Li et al., 2023) by over 3%. Figure 3 provides a training example curated by TrainRef with distributional label in the instance-dependent 40% noise setting on CIFAR-100. The image belongs to the class “sea” but has been incorrectly labeled as “plain”. TrainRef assigns a distributional label that balances between

Table 1: Comparison with SOTA methods on CIFAR-100 datasets with different types and levels of label noise. Mean  $\pm$  standard deviation is reported over 3 runs. The results are primarily derived from (Li et al., 2023) or the original papers.

Noise type	Sym 20%	Sym 50%	Sym 80%	Asym 40%	Inst 40%
CE	55.17 $\pm$ 0.12	32.40 $\pm$ 0.16	7.70 $\pm$ 0.16	40.63 $\pm$ 0.26	43.17 $\pm$ 0.15
Decoupling (Malach & Shalev-Shwartz, 2017)	52.75 $\pm$ 0.11	27.59 $\pm$ 0.16	7.38 $\pm$ 0.09	39.12 $\pm$ 0.08	-
Co-teaching (Han et al., 2018)	51.24 $\pm$ 0.23	25.07 $\pm$ 0.18	8.50 $\pm$ 0.06	38.06 $\pm$ 0.15	23.21 $\pm$ 0.57
JointOptim (Tanaka et al., 2018)	58.50 $\pm$ 0.47	53.58 $\pm$ 0.43	24.62 $\pm$ 0.50	61.17 $\pm$ 0.39	-
Co-teaching+ (Yu et al., 2019)	51.24 $\pm$ 0.23	25.07 $\pm$ 0.18	9.50 $\pm$ 0.08	36.58 $\pm$ 0.16	24.45 $\pm$ 0.71
GCE (Zhang & Sabuncu, 2018)	76.16 $\pm$ 0.11	72.84 $\pm$ 0.12	28.40 $\pm$ 0.06	46.08 $\pm$ 0.20	45.69 $\pm$ 0.14
PENCIL (Yi & Wu, 2019)	55.17 $\pm$ 0.12	37.12 $\pm$ 0.17	9.33 $\pm$ 0.33	40.63 $\pm$ 0.26	-
JoCoR (Wei et al., 2020)	54.70 $\pm$ 0.08	26.45 $\pm$ 0.13	7.35 $\pm$ 0.05	37.09 $\pm$ 0.09	23.95 $\pm$ 0.44
DivideMix (Li et al., 2020)	76.16 $\pm$ 0.11	72.84 $\pm$ 0.12	28.40 $\pm$ 0.06	55.56 $\pm$ 0.53	76.08 $\pm$ 0.35
ELR (Liu et al., 2020)	69.93 $\pm$ 0.14	58.10 $\pm$ 0.17	28.40 $\pm$ 0.06	46.08 $\pm$ 0.20	-
ELR+ (Liu et al., 2020)	76.94 $\pm$ 0.18	73.01 $\pm$ 0.14	58.01 $\pm$ 0.17	74.39 $\pm$ 0.17	-
Co-learning (Wei et al., 2020)	69.93 $\pm$ 0.14	58.10 $\pm$ 0.17	41.77 $\pm$ 0.32	51.50 $\pm$ 0.24	-
DISC (Li et al., 2023)	78.75 $\pm$ 0.13	75.21 $\pm$ 0.15	57.61 $\pm$ 0.29	76.50 $\pm$ 0.15	78.44 $\pm$ 0.19
L2B-C2D (Zhou et al., 2024)	79.67 $\pm$ 0.14	78.23 $\pm$ 0.16	69.66 $\pm$ 0.19	78.22 $\pm$ 0.14	79.43 $\pm$ 0.17
<b>Ours</b>	<b>85.44 <math>\pm</math> 0.21</b>	<b>82.07 <math>\pm</math> 0.17</b>	<b>77.85 <math>\pm</math> 0.35</b>	<b>79.67 <math>\pm</math> 0.22</b>	<b>82.33 <math>\pm</math> 0.16</b>

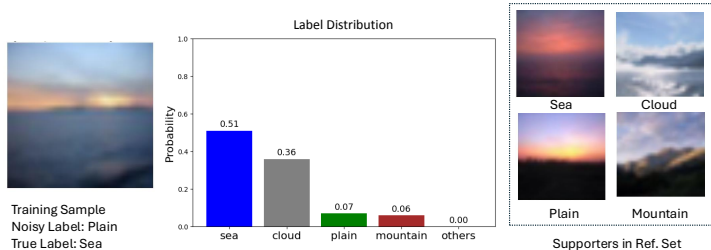


Figure 3: A noisy training sample (in the left) is curated by TrainRef to have a class distribution where the probability over the class “sea”, “cloud”, “plain”, “mountain”, and etc (in the middle). In addition, the reference samples voting for such a curation is shown in the right.

“sea” and “cloud”, reflecting the inherent ambiguity in the image. In addition, we show reference samples voting this label, explaining how TrainRef makes such a curation decision. More examples are available at our website (TrainRef, 2025).

**Results on WebVision and Animal-10N.** Tables 2 and Tables 3 present the generalization performance of TrainRef with SOTA methods on real-world noisy datasets, WebVision and Animal-10N, respectively. The results demonstrate that TrainRef improves the predictive accuracy by 2% - 4% on both benchmarks, highlighting its effectiveness in handling real-world label noise.

## 4.2 RQ2: CONFIDENCE RELIABILITY

**Setup.** We evaluate calibration on CIFAR-100 under three noise regimes: noise-free, symmetric 20%, and symmetric 80%. Baseline calibration methods include (i) *Raw cross-entropy (CE) loss*, (ii) *Focal variants loss* (Focal, AdaFocal, DualFocal/AdaDualFocal (Mukhoti et al., 2020a; Ghosh et al., 2022; Tao et al., 2023)), (iii) *Post-hoc calibration* such as Temperature Scaling (TS) (Guo et al., 2017), PTS (Tomani et al., 2022), Spline (Gupta et al., 2020), MnM (Zhang et al., 2020)), (iv) *Denoising methods* (DISC, L2B) with and without TS, and TrainRef with and without TS. We report accuracy, ECE (Nixon et al., 2019), and AdaECE (Mukhoti et al., 2020b). Extended results with additional metrics ( $ECE_{\text{debias}}$ ,  $ECE_{\text{sweep}}$  (Roelofs et al., 2022)), other CIFAR-100 noise settings, and two real-world datasets (Animal-10N, WebVision) are provided in the Appendix D.

**Results.** Table 4 shows that TrainRef consistently achieves the best trade-off between accuracy and calibration across all noise settings. Focal variants losses amplify noise by up-weighting mislabeled “hard” samples, while post-hoc methods such as TS only rescale logits  $\hat{p}_k = \text{softmax}(z_k/T)$  and cannot fix incorrect confidence rankings, yielding low ECE but poor accuracy. Denoising methods such as DISC and L2B improve accuracy through filtering or hard relabeling, but discard uncertain yet informative samples, thereby weakening calibration. By contrast, TrainRef utilizes soft labels

Table 2: Prediction Accuracy on WebVision.

Accuracy (%)	Top-1	Top-5
F-correction (Patrini et al., 2017)	61.12	82.68
MentorNet (Jiang et al., 2018)	63.00	81.40
Co-teaching (Han et al., 2018)	63.58	85.20
ELR (Liu et al., 2020)	76.26	91.26
DivideMix (Li et al., 2020)	77.32	91.64
ELR+ (Liu et al., 2020)	77.78	91.68
GJS (Engleson & Azizpour, 2021)	77.99	90.62
CC (Zhao et al., 2022)	79.36	93.64
DISC (Li et al., 2023)	80.28	92.28
LSL (Kim et al., 2024)	81.40	93.00
<b>Ours</b>	<b>82.28</b>	<b>95.14</b>
<b>Ours (IN1k Pretrained)</b>	<b>84.10</b>	<b>96.34</b>

Table 3: Prediction Accuracy on Animal-10N.

Method	Accuracy (%)
CE (Engleson & Azizpour, 2021)	79.4 ± 0.14
GCE (Zhang & Sabuncu, 2018)	81.5 ± 0.08
SELFIE (Song et al., 2019)	81.8 ± 0.09
Mixup (Zhang, 2017)	82.7 ± 0.03
Co-learning (Tan et al., 2021)	83.0
PLC (Zhang et al., 2021)	83.4 ± 0.43
Nested Co-teaching (Chen et al., 2021b)	84.1 ± 0.10
GJS (Engleson & Azizpour, 2021)	84.2 ± 0.07
DISC (Li et al., 2023)	87.1 ± 0.15
LSL (Kim et al., 2024)	89.1
<b>Ours</b>	<b>90.90 ± 0.24</b>
<b>Ours (IN1k Pretrained)</b>	<b>93.72 ± 0.15</b>

Table 4: Comparison of confidence calibration performance across noise-free, symmetric 20% (Sym20), and symmetric 80% (Sym80) settings on CIFAR-100. Results report test accuracy (higher is better) and calibration errors (ECE, AdaECE; lower is better). Grey-shaded rows indicate methods where temperature scaling (TS) is applied on top of the base method. Bold entries mark the best results under the same TS setting (either with TS or without TS).

Method	Noise-Free			Sym20			Sym80		
	Test Acc (%)	ECE (↓)	AdaECE (↓)	Test Acc (%)	ECE (↓)	AdaECE (↓)	Test Acc (%)	ECE (↓)	AdaECE (↓)
CE (Baseline)	77.87	0.1512	0.1508	51.76	0.088	0.0879	16.38	0.0946	0.0946
Runtime Method									
Focal Loss (Mukhoti et al., 2020a)	78.31	0.0864	0.0866	52.16	0.1199	0.1198	16.26	0.1055	0.1055
Ada Focal Loss (Ghosh et al., 2022)	78.55	0.0723	0.0717	51.69	0.0923	0.0913	16.68	0.105	0.1049
Dual Focal Loss (Tao et al., 2023)	77.93	0.0925	0.0924	47.32	0.1476	0.1476	16.95	0.1057	0.1055
CE + Posthoc									
CE+TS (Guo et al., 2017)	77.87	0.0293	0.0297	51.76	<b>0.0137</b>	<b>0.0138</b>	16.38	0.0136	0.0097
CE+PTS (Tomani et al., 2022)	77.87	0.0254	0.0266	51.76	0.0263	0.028	16.38	0.014	0.0135
CE+Spline (Gupta et al., 2020)	77.87	0.0306	0.0331	51.76	0.0242	0.028	16.38	0.024	0.0286
CE+MnM (Zhang et al., 2020)	77.87	0.0212	0.0201	51.76	0.0177	0.0126	16.38	0.0134	<b>0.0085</b>
Mixture									
DISC (Li et al., 2023)	81.23 ± 0.10	0.113 ± 0.013	0.112 ± 0.011	78.75 ± 0.13	0.118 ± 0.011	0.114 ± 0.016	57.61 ± 0.29	0.12 ± 0.013	0.147 ± 0.016
DISC+TS	81.23 ± 0.10	0.025 ± 0.007	0.027 ± 0.007	78.75 ± 0.13	0.043 ± 0.005	0.045 ± 0.010	57.61 ± 0.29	0.061 ± 0.007	0.053 ± 0.012
L2B (Zhou et al., 2024)	82.31 ± 0.14	0.124 ± 0.011	0.131 ± 0.009	79.67 ± 0.14	0.103 ± 0.013	0.112 ± 0.009	69.66 ± 0.19	0.133 ± 0.009	0.152 ± 0.022
L2B+TS	82.31 ± 0.14	0.027 ± 0.008	0.028 ± 0.009	79.67 ± 0.14	0.042 ± 0.012	0.043 ± 0.011	69.66 ± 0.19	0.057 ± 0.015	0.061 ± 0.017
Ours	<b>85.87 ± 0.15</b>	<b>0.041 ± 0.008</b>	<b>0.043 ± 0.010</b>	<b>85.44 ± 0.21</b>	<b>0.048 ± 0.009</b>	<b>0.047 ± 0.008</b>	<b>77.85 ± 0.35</b>	<b>0.082 ± 0.013</b>	<b>0.086 ± 0.011</b>
Ours+TS	<b>85.87 ± 0.15</b>	<b>0.015 ± 0.007</b>	<b>0.014 ± 0.008</b>	<b>85.44 ± 0.21</b>	0.015 ± 0.009	0.016 ± 0.006	<b>77.85 ± 0.35</b>	<b>0.011 ± 0.005</b>	0.014 ± 0.009

and a minimal reference set  $\mathcal{D}_{ref}$  to retain uncertainty and provide reliable supervision, resulting in stronger calibration and accuracy.

### 4.3 RQ3: USER STUDY

**Setup.** We collect confidence scores from the test set after training the models on different noise levels (i.e., 20%, 50%, and 80%) of CIFAR-100. Specifically, under a noisy rate (e.g., 20%), we learn TrainRef and the baseline DISC. Then, we select 100 test samples (1) which are predicted to have low confidence by either TrainRef or DISC; or (2) where TrainRef and DISC have a large disagreement on their confidence. We hire 5 experts as participants, each with over 3 years of experience in model training and data labeling. Each participant is presented with two anonymous predictions (TrainRef or the baseline) and asked to choose a predicted confidence to agree with.

**Results.** Table 5 shows that across all noise settings, participants consistently prefer the predictions from TrainRef by a significant margin. As the noise rate increases, the preference for TrainRef becomes even more pronounced. Notably, at an 80% noise rate, in over 75% of cases, participants consider TrainRef’s predictions to be more reliable, showing its advantage to produce semantically meaningful and robust classifications under extreme noise conditions. More examples and results are available at our website (TrainRef, 2025).

### 4.4 RQ4: ABLATION STUDY

**Effectiveness of Initial Reference Set Size.** Table 6 shows the effect of varying the initial reference set size  $|\mathcal{D}_{ref}|$  on the noise rate of the augmented set  $\mathcal{D}_{ref}^*$  and test accuracy. Results are reported on CIFAR-100 with 80% symmetric noise and on two real-world datasets (Animals-10N and WebVision), for which only test accuracy is available due to the lack of clean labels. TrainRef consistently maintains a low noise rate (< 5%) in the augmented reference set and high test accuracy across varying

Table 5: User Study Evaluation of Prediction Reliability: At each noise rate, the participants choose to agree with the predicted confidence of the test sample by either TrainRef or DISC.

Sym. Noise Rate	TrainRef (%)	DISC (%)	Total (%)
20%	62%	38%	100%
50%	74%	26%	100%
80%	78%	22%	100%

Table 6: Noise Rate and Sample Counts vs Initial Reference Set Size

Dataset & Noise	Metric	k=1	k=5	k=10	k=100
CIFAR-100 (Sym 80%)	CleanNum	6069	6352	6389	6409
	Mis	140	110	103	101
	NR( $\downarrow$ )	0.0230	0.0173	0.0161	0.0157
	Test Acc	77.85	77.89	77.93	77.91
WebVision	Test Acc	82.21	82.28	82.27	82.36
Animals-10N	Test Acc	90.75	90.90	90.97	90.88

Table 7: Effectiveness of Label Distribution on Calibration (CIFAR100 Inst. 40%)

Method	Hard Relabel	Soft Relabel	NNL $\downarrow$	ECE $\downarrow$	Test Acc (%)
Ours			0.710	0.055	81.07
Ours-TC			0.707	0.032	–
Ours	✓		0.756	0.065	81.77
Ours-TC	✓		0.740	0.032	–
Ours	✓	✓	<b>0.683</b>	<b>0.046</b>	<b>82.33</b>
Ours-TC	✓	✓	<b>0.669</b>	<b>0.017</b>	–

reference sizes, remaining effective even with just one reference sample per class ( $|D_{ref}| = 1$ ). In our experiments, we use  $|D_{ref}| = 5$  to balance accessibility and performance. These results demonstrate TrainRef’s ability to efficiently identify clean samples under extreme noise and minimal supervision. **Effectiveness of Label Distribution** We evaluate TrainRef on CIFAR-100 with 40% instance-dependent noise, focusing on the effect of label distribution. Specifically, we replace soft labels with one-hot labels and apply temperature scaling as in RQ2. Lower ECE (Expected Calibration Error) and NNL (Negative Log-likelihood) indicate better calibration. As shown in Table 7, TrainRef achieves the lowest ECE with and without temperature scaling, outperforming all ablations. While one-hot labels (hard relabel) yield similar test accuracy, they degrade calibration by sharpening decision boundaries and ignoring ambiguous samples.

Additional ablations on Phase III iteration counts (Appendix F.2), the role of influence-based reference augmentation (Appendix F.1), as well as analyses of computational cost (Appendix B), backbone fairness (Appendix F.5) and limitations (Appendix H) are provided in the Appendix.

## 5 RELATED WORK

**Learning with Noisy Labels (& Data Curation)** Early LNL methods (Hendrycks et al., 2018; Patrini et al., 2017) assume class-conditional noise modeled by a label transition matrix. However, theoretical work (Chen et al., 2021a; Xia et al., 2020) shows that real-world noise is largely instance-dependent, making transition matrix estimation both inaccurate and computationally expensive. To handle instance-specific noise, prior work proposes re-weighting or filtering noisy samples using loss, confidence, or multi-view signals (Han et al., 2018; Yu et al., 2019; Li et al., 2020; Zhu et al., 2021; Kim et al., 2024). Semi-supervised learning (SSL) approaches (Sohn et al., 2020; Li et al., 2023) assign pseudo-labels to noisy samples, often relying on augmentations (Nishi et al., 2021; Cubuk et al., 2020). However, these heuristics degrade under high noise, mislabeling ambiguous samples and harming generalization (Das & Sanghavi, 2023). Meta-learning methods (Wu et al., 2021; Li et al., 2019) use clean references for guidance but incur high computational cost due to bi-level optimization.

**Confidence Calibration.** Confidence calibration aims to align predicted probabilities with true correctness likelihoods. Classical post-hoc methods include Temperature Scaling (TS) (Guo et al., 2017), parameterized transformations such as PTS (Tomani et al., 2022), spline-based mappings (Gupta et al., 2020), and Mix-n-Match (MnM) (Zhang et al., 2020). These methods adjust output probabilities after training, but they cannot fix mis-ordered confidence rankings learned under noise. Train-time calibration has also been explored: focal-style losses (Mukhoti et al., 2020a; Ghosh et al., 2022; Tao et al., 2023) emphasize hard samples but risk amplifying label noise, while denoising approaches such as DISC (Li et al., 2023) and L2B (Zhou et al., 2024) combine label correction with calibration. Evaluation metrics like ECE (Nixon et al., 2019), AdaECE (Mukhoti et al., 2020a), and recent unbiased estimators such as  $ECE_{debias}$  and  $ECE_{sweep}$  (Roelofs et al., 2022) provide multiple perspectives on calibration quality. Our work builds upon these foundations by integrating calibration with noise-robust training via reference-guided distributional curation.

## 6 CONCLUSION

We propose TrainRef, a training-time data curation framework that unifies label denoising and confidence calibration. Through Curation-Training Co-evolution, TrainRef refines the embedding space, maintains a diverse reference set, and assigns reliable soft labels. Extensive experiments show it outperforms state-of-the-art methods, improving accuracy and confidence calibration, with qualitative studies confirming its reliability for real-world noisy-label scenarios. In the future, we will deliver a library of TrainRef for the community and generalize the technique on generative models.

### ACKNOWLEDGMENTS

We would like to thank Koh Pang Wei from the University of Washington for his insightful discussions and inspiration. This research is supported in part by the National Natural Science Foundation of China (62572300), the Ministry of Education, Singapore (MOE-T2EP20124-0017, MOET32020-0004), the National Research Foundation, Singapore and the Cyber Security Agency under its National Cybersecurity R&D Programme (NCRP25-P04-TAICeN), DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008-1B), and the Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

### REFERENCES

- CVAT AI. How much does it cost to outsource annotation to a data labeling service? <https://www.cvat.ai/blog/how-much-does-it-cost-to-outsource-annotation-to-a-data-labeling-service>, 2024.
- 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*, volume 35, pp. 11442–11450, 2021a.
- Yingyi Chen, Xi Shen, Shell Xu Hu, and Johan AK Suykens. Boosting co-teaching with compression regularization for label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2688–2692, 2021b.
- Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/757f843a169cc678064d9530d12a1881-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/757f843a169cc678064d9530d12a1881-Paper.pdf).
- Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702–703, 2020.
- Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In *International Conference on Machine Learning*, pp. 7102–7140. PMLR, 2023.
- Elahe Delavari, Aws Khalil, and Jaerock Kwon. Caril: Confidence-aware regression in imitation learning for autonomous driving, 2025. URL <https://arxiv.org/abs/2503.00783>.
- Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1): 1–22, 1977.
- Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.
- Forbes. Andrew ng launches a campaign for data-centric ai. <https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/>, 2024.

- Arindam Ghosh, Thomas Schaaf, and Matthew R. Gormley. Adafocal: Calibration-aware adaptive focal loss. In *Advances in Neural Information Processing Systems*, 2022.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.
- Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. *arXiv preprint arXiv:2006.12800*, 2020.
- Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.
- Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018.
- Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.
- Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*, 2025.
- Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9676–9686, 2022.
- Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, and Hwanjo Yu. Learning discriminative dynamics with label corruption for noisy label detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22477–22487, 2024.
- Youngdong Kim, Juseung Yun, Hyounguk Shon, and Junmo Kim. Joint negative and positive learning for noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9442–9451, 2021.
- Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.
- Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- Jichang Li, Guanbin Li, Feng Liu, and Yizhou Yu. Neighborhood collective estimation for noisy label identification and correction. In *European Conference on Computer Vision*, pp. 128–145. Springer, 2022.
- Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5051–5059, 2019.
- Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.
- Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.
- Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24070–24079, 2023.

- Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.
- Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pp. 14153–14172. PMLR, 2022.
- Yangdi Lu and Wenbo He. Selc: self-ensemble label correction improves learning with noisy labels. *arXiv preprint arXiv:2205.01156*, 2022.
- Yangdi Lu, Yang Bo, and Wenbo He. Co-matching: Combating noisy labels by augmentation anchoring. *arXiv preprint arXiv:2103.12814*, 2021.
- Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". *Advances in neural information processing systems*, 30, 2017.
- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, pp. 15744–15755, 2020a.
- Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H.S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems*, 2020b. URL <https://proceedings.neurips.cc/paper/2020/file/aeb7b30ef1d024a76f21a1d40e30c302-Paper.pdf>.
- Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.
- Kento Nishi, Yi Ding, Alex Rich, and Tobias Hollerer. Augmentation strategies for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8022–8031, 2021.
- Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 38–41, 2019.
- Feiyang Pan, Jia He, Dandan Tu, and Qing He. Trust the model when it is confident: Masked model-based actor-critic. *Advances in neural information processing systems*, 33:10537–10546, 2020.
- Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2536–2544, 2016.
- 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*, pp. 1944–1952, 2017.
- Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.
- Lorenzo Perini and Jesse Davis. Unsupervised anomaly detection with rejection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 69673–69691. Curran Associates, Inc., 2023. URL [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/dc48c738d3ef8c81b6e968453a84a819-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/dc48c738d3ef8c81b6e968453a84a819-Paper-Conference.pdf).
- Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020.

- Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C. Mozer. Mitigating bias in calibration error estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*. PMLR, 2022.
- Rikard Rosenbacke, Åsa Melhus, and David Stuckler. False conflict and false confirmation errors are crucial components of ai accuracy in medical decision making. *Nature Communications*, 15(1): 6896, 2024.
- Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.
- Mengmeng Sheng, Zeren Sun, Tao Chen, Shuchao Pang, Yucheng Wang, and Yazhou Yao. Foster adaptivity and balance in learning with noisy labels. In *European Conference on Computer Vision*, pp. 217–235. Springer, 2024.
- Xiaoshuang Shi, Zhenhua Guo, Kang Li, Yun Liang, and Xiaofeng Zhu. Self-paced resistance learning against overfitting on noisy labels. *Pattern Recognition*, 134:109080, 2023.
- Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pp. 5907–5915. PMLR, 2019.
- Zeren Sun, Huafeng Liu, Qiong Wang, Tianfei Zhou, Qi Wu, and Zhenmin Tang. Co-ldl: A co-training-based label distribution learning method for tackling label noise. *IEEE Transactions on Multimedia*, 24:1093–1104, 2021.
- Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 1405–1413, 2021.
- Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5552–5560, 2018.
- Linwei Tao, Minjing Dong, and Chang Xu. Dual focal loss for calibration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.
- Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European conference on computer vision*, pp. 555–569. Springer, 2022.
- TrainRef. Trainref, 2025. <https://sites.google.com/view/train-ref/home>.
- Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13726–13735, 2020.
- Yichen Wu, Jun Shu, Qi Xie, Qian Zhao, and Deyu Meng. Learning to purify noisy labels via meta soft label corrector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10388–10396, 2021.
- 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*, 33:7597–7610, 2020.
- Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5192–5201, 2021.

- Kun Yi and Jianxin Wu. 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.
- Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International conference on machine learning*, pp. 7164–7173. PMLR, 2019.
- Hongyi Zhang. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.
- Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pp. 11117–11128. PMLR, 2020.
- Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. *arXiv preprint arXiv:2103.07756*, 2021.
- Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.
- Ganlong Zhao, Guanbin Li, Yipeng Qin, Feng Liu, and Yizhou Yu. Centrality and consistency: two-stage clean samples identification for learning with instance-dependent noisy labels. In *European Conference on Computer Vision*, pp. 21–37. Springer, 2022.
- Evgenii Zheltonozhskii, Chaim Baskin, Avi Mendelson, Alex M Bronstein, and Or Litany. Contrast to divide: Self-supervised pre-training for learning with noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1657–1667, 2022.
- Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Asymmetric loss functions for noise-tolerant learning: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8094–8109, 2023.
- Yuyin Zhou, Xianhang Li, Fengze Liu, Qingyue Wei, Xuxi Chen, Lequan Yu, Cihang Xie, Matthew P Lungren, and Lei Xing. L2b: Learning to bootstrap robust models for combating label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23523–23533, 2024.
- Zhaowei Zhu, Tongliang Liu, and Yang Liu. 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.